Universidade de Aveiro
2018

Departamento de
Eletrónica, Telecomunicações e Informática
Programa de Doutoramento em Informática das
Universidades do Minho, Aveiro e do Porto

Pedro Miguel
Lopes Sernadela

# Serviços de Integração de Dados para Aplicações Biomédicas

# Data Integration Services for Biomedical Applications

Universidade de Aveiro
2018

Departamento de
Eletrónica, Telecomunicações e Informática
Programa de Doutoramento em Informática das
Universidades do Minho, Aveiro e do Porto

Pedro Miguel
Lopes Sernadela

# Serviços de Integração de Dados para Aplicações Biomédicas

# Data Integration Services for Biomedical Applications

Tese apresentada às Universidades de Aveiro, Minho e Porto para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Informática (MAP-i), realizada sob a orientação científica do Doutor José Luís Guimarães Oliveira, Professor Associado com Agregação do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.

**o júri / the jury**

presidente / president

**Doutor Artur Manuel Soares da Silva**
Professor Catedrático da Universidade de Aveiro

vogais / examiners committee

**Doutor Rui Pedro Sanches de Castro Lopes**
Professor Coordenador da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Bragança

**Doutor Paulo José Osório Rupino da Cunha**
Professor Auxiliar com Agregação do Departamento de Engenharia Informática da Universidade Coimbra

**Doutor Francisco José Moreira Couto**
Professor Associado com Agregação do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa

**Doutor Luís Manuel Dias Coelho Soares Barbosa**
Professor Associado com Agregação do Departamento de Informática da Universidade do Minho

**Doutor José Luís Guimarães Oliveira**
Professor Associado com Agregação do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro (orientador)

**agradecimentos**

Gostaria de expressar o meu especial agradecimento ao grupo de bioinformática do Instituto de Engenharia Eletrónica e Informática de Aveiro (IEETA) pelo ótimo espaço de cooperação e ambiente de trabalho proporcionado pelos colegas. Em particular, um muito obrigado ao meu orientador prof. José Luís Oliveira pela oportunidade, partilha das suas visões, recomendações e todo o processo de orientação que foram fundamentais para a execução deste trabalho. Por último, mas não menos importante, gostaria de agradecer à minha família e amigos pelo apoio, amizade e paciência. Finalmente, um grande agradecimento à Fundação para a Ciência e Tecnologia (FCT) por tornar possível este doutoramento, através da bolsa SFRH/BD/52484/2014.

**acknowledgements**

I would like to express my sincere appreciation to the bioinformatics group at "Instituto de Engenharia Eletrónica e Informática de Aveiro" (IEETA) for the great working environment provided, the productive discussions and overall cooperation. In this regard, special thanks to prof. José Luís Oliveira for giving me the opportunity to carry out my research project, for his guidance and unconditional support. Last but not least, I want to extend my deepest gratitude to my family and friends for their friendship, support and patience. Finally, I gratefully acknowledge the "Fundação para a Ciência e Tecnologia" (FCT) for making this Ph.D. work possible, through grant SFRH/BD/52484/2014.

**Palavras-chave** Web semântica, integração de dados, extração de informação, bioinformática, bases de conhecimento, interoperabilidade, doenças raras.

**Resumo** Nas últimas décadas, o campo das ciências biomédicas proporcionou grandes avanços científicos estimulados pela constante evolução das tecnologias de informação. A criação de diversas ferramentas na área da bioinformática e a falta de integração entre novas soluções resultou em enormes quantidades de dados distribuídos por diferentes plataformas. Dados de diferentes tipos e formatos são gerados e armazenados em vários repositórios, o que origina problemas de interoperabilidade e atrasa a investigação. A partilha de informação e o acesso integrado a esses recursos são características fundamentais para a extração bem sucedida do conhecimento científico.

Nesta medida, esta tese fornece contribuições para acelerar a integração, ligação e reutilização semântica de dados biomédicos. A primeira contribuição aborda a interconexão de registos distribuídos e heterogéneos. A metodologia proposta cria uma visão holística sobre os diferentes registos, suportando a representação semântica de dados e o acesso integrado. A segunda contribuição aborda a integração de diversos dados para investigações científicas, com o objetivo de suportar serviços interoperáveis para a partilha de informação. O terceiro contributo apresenta uma arquitetura modular que apoia a extração e integração de informações textuais, permitindo a exploração destes dados. A última contribuição consiste numa plataforma web para acelerar a criação de sistemas de informação semânticos. Todas as soluções propostas foram validadas no âmbito das doenças raras.

**Keywords** Semantic web, data integration, information extraction, bioinformatics, knowledge bases, interoperability, rare diseases.

**Abstract** In the last decades, the field of biomedical science has fostered unprecedented scientific advances. Research is stimulated by the constant evolution of information technology, delivering novel and diverse bioinformatics tools. Nevertheless, the proliferation of new and disconnected solutions has resulted in massive amounts of resources spread over heterogeneous and distributed platforms. Distinct data types and formats are generated and stored in miscellaneous repositories posing data interoperability challenges and delays in discoveries. Data sharing and integrated access to these resources are key features for successful knowledge extraction.

In this context, this thesis makes contributions towards accelerating the semantic integration, linkage and reuse of biomedical resources. The first contribution addresses the connection of distributed and heterogeneous registries. The proposed methodology creates a holistic view over the different registries, supporting semantic data representation, integrated access and querying. The second contribution addresses the integration of heterogeneous information across scientific research, aiming to enable adequate data-sharing services. The third contribution presents a modular architecture to support the extraction and integration of textual information, enabling the full exploitation of curated data. The last contribution lies in providing a platform to accelerate the deployment of enhanced semantic information systems. All the proposed solutions were deployed and validated in the scope of rare diseases.

# List of contents

# List of figures

# List of tables

# List of acronyms

| | |
|---|---|
| **AO** | Annotation Ontology |
| **API** | Application Programming Interface |
| **ChEBI** | Chemical Entities of Biological Interest |
| **CL** | Cell Ontology |
| **CRAFT** | Colorado Richly Annotated Full-Text |
| **CRF** | Conditional Random Field |
| **CSV** | Comma-Separated Values |
| **CTD** | Comparative Toxicogenomics Database |
| **CUI** | Concept Unique Identifier |
| **DICOM** | Digital Imaging and Communications in Medicine |
| **DINTO** | Drug-Drug Interactions Ontology |
| **DMD** | Duchenne Muscular Dystrophy |
| **DO** | Disease Ontology |
| **EHR** | Electronic Health Record |
| **ETL** | Extract-Transform-and-Load |
| **ExPASy** | Expert Protein Analysis System |
| **FOAF** | Friend Of A Friend |
| **FTP** | File Transfer Protocol |
| **GeneRIF** | Gene Reference Into Function |
| **GO** | Gene Ontology |
| **GOA** | GO Annotation |
| **GREC** | Gene Regulation Event Corpus |
| **GUI** | Graphical User Interface |

| | |
|---|---|
| **HGNC** | HUGO Gene Nomenclature Committee |
| **HGP** | Human Genome Project |
| **HMDB** | Human Metabolome Database |
| **HPO** | Human Phenotype Ontology |
| **HTML** | Hypertext Markup Language |
| **HTTP** | Hypertext Transfer Protocol |
| **HUGO** | Human Genome Organisation |
| **IE** | Information Extraction |
| **IPI** | International Protein Index |
| **IRDiRC** | International Rare Diseases Research |
| **JSON** | JavaScript Object Notation |
| **KEGG** | Kyoto Encyclopedia of Genes and Genomes |
| **LOD** | Linked Open Data |
| **MedDRA** | Medical Dictionary for Regulatory Activities |
| **MEDLARS** | Medical Literature Analysis and Retrieval System |
| **MEDLINE** | MEDLARS Online |
| **MeSH** | Medical Subject Headings |
| **ML** | Machine Learning |
| **NAR** | Nucleic Acids Research |
| **NCBI** | National Center for Biotechnology Information |
| **NER** | Named Entity Recognition |
| **NGS** | Next-Generation Sequencing |
| **NLM** | National Library of Medicine |
| **NLP** | Natural Language Processing |
| **OMIM** | Online Mendelian Inheritance in Man |
| **ORM** | Object Relation Mapping |
| **OWL** | Web Ontology Language |
| **PDB** | Protein Data Bank |
| **PharmGKB** | Pharmacogenomics Knowledge Base |
| **PoS** | Part-of-Speech |

| | |
|---|---|
| **PRO** | Protein Ontology |
| **RDBMS** | Relational Database Management System |
| **RDF** | Resource Description Framework |
| **RDFS** | Resource Description Framework (RDF) Schema |
| **REST** | Representational State Transfer |
| **RO** | Research Objects |
| **SNOMED** | Systematized Nomenclature of Medicine |
| **SO** | Sequence Ontology |
| **SPARQL** | SPARQL Protocol and RDF Query Language |
| **SQL** | Structured Query Language |
| **SW** | Semantic Web |
| **TDB** | Transactional Database |
| **TOXNET** | Toxicology Data Network |
| **UMLS** | Unified Medical Language System |
| **UniProt** | Universal Protein Resource |
| **URI** | Uniform Resource Identifier |
| **W3C** | World Wide Web Consortium |
| **WSD** | Word Sense Disambiguation |
| **WWW** | World Wide Web |
| **XML** | Extensible Markup Language |
| **XPath** | XML Path Language |

# Chapter 1

# Introduction

Bioinformatics has been one the most active areas of computer science. Since the Human Genome Project (HGP) revolution to decode human genetic code [1], the union of life and computer sciences has fostered unprecedented advances in several multidisciplinary areas. Sequence decoding, genetic studies and drug advances are just sample areas where the efficient collaboration between computer scientists and biologists has been beneficial for scientific innovation.

Over the years, this successful partnership has introduced great modifications on how researchers access and use computation tools for scientific discovery. These changes are so significant, that today it is almost unfeasible to conceive a successful biomedicine project without computational technology. This setting directly implies that bioinformatics advances are highly dependent on computational technology innovation, being crucial to explore novel software and hardware tools.

Nowadays, that strong relationship is still driving scientific research to a higher level of automation, increasingly stimulating the development of specialized tools and revolutionizing all biomedical research fields. The outcome of this revolution has been an infinite number of computer-based resources to deal with. Indeed, many services, databases, systems and applications have attempted to solve existing research issues. Although the increase of such resources appears to be logically beneficial, the effects of having such a variety are questionable. The cost of maintaining this research field is high due to non-integrated software with too many competing models and architectures. Making the situation even more challenging, biomedical research has yet to deal with the disparity of user and data interfaces. Different formats, styles or contents are an issue for the solutions developed, hindering resource linkage and delaying the advance of research

on life sciences.

With these challenges in mind, this research effort introduces newly interoperable solutions to keep bioinformatics applications at the boundaries of computer science innovation.

## 1.1 Motivation

The latest biomedicine advances have brought exceptional changes in how biomedical resources are handled. Extensive and large-scale scientific discovery methodologies revolutionized the way that bio-resources are studied, moving most scientific efforts into data-intensive science-oriented research. For instance, the vision of studying datasets individually loses strength towards a fresh paradigm of assessing them as a global and connectable structure through the many different fields of biomedicine: genomics, proteomics, metabolomics, pharmacogenomics, among others. Although this provides a new way of analyzing the whole biological spectrum, it rapidly led to an exponential increase in the amount of data and repositories to explore over different levels. Exploring this huge amount of resources consumes expensive and valuable resources, both human and technical, and acquiring new insights into the existing disconnected knowledge is challenging.

On the one hand, unstructured information, such as articles, books and technical reports are challenging to analyze. At the same time, high amounts of data are increasing day by day. The MEDLINE database, for instance, shows continuous annual growth [2], and in 2015 contained a total of 23 million references to journal articles in fields related to the life sciences (Figure 1.1).

The lack of more concise knowledge makes it harder than ever for researchers to find and assimilate all the information relevant to their research. For instance, extracting biomedical associations or hypotheses from narrative documents is not a trivial task and requires highly trained data curators that create and update annotation resources, a time-consuming and expensive task. This situation triggered various research efforts trying to automate and summarize the knowledge scattered across multiple publications and store it in a structured form. Regarding the biomedical domain, significant progress has been made in the use of computerized solutions to aid in the analysis, extraction and storage of relevant concepts, and their respective attributes and relationships (Figure 1.2).

Although full natural language understanding is far from complete through these

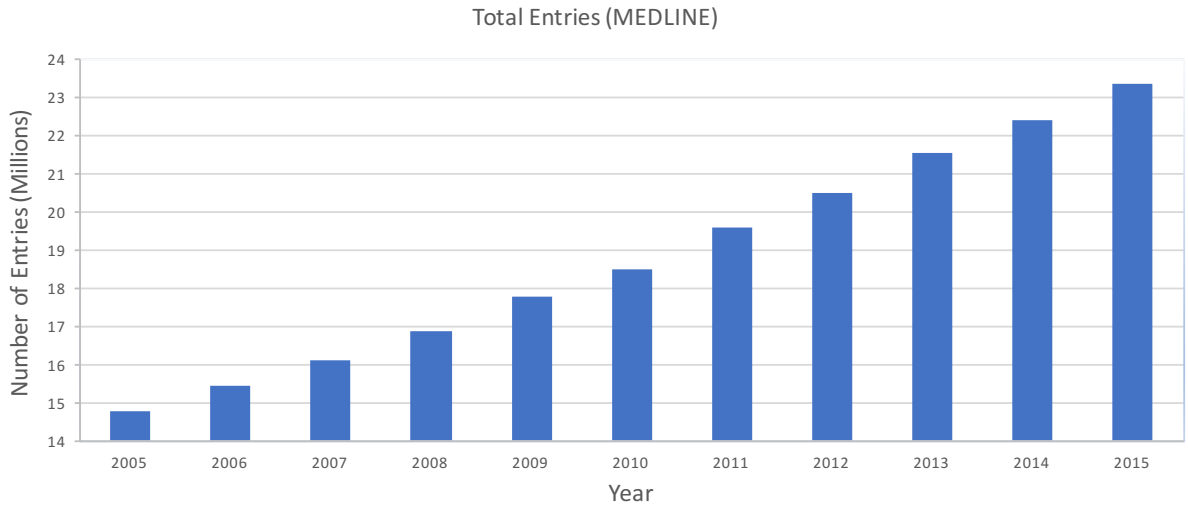


Figure 1.1: MEDLINE bibliographic database growth.

solutions, they provide good assistance to assimilate the high rate of new publications [3], and to discover indirect associations between important biomedical elements. These computerized solutions have been increasingly applied to assist bio-curators, allowing the extraction of relevant information regarding biomedical concepts such as genes, proteins, chemical compounds or diseases [4], and thus reducing curation times and cost [5].

Several text-mining algorithms and technologies have appeared to offer more effective mining solutions, exploring complex relation extraction processes, life-science terminology normalization and information fusion [6]. These challenges transcend many disciplines, which makes it more difficult to find and integrate all the relevant information from different research communities. Nonetheless, these efforts are still hindered by a lack of standardized ways to process the vast amount of data generated [7]. This concern can be split into two major challenges. First, there are interoperability issues between information extraction components for concept recognition and relation extraction methodologies. Second, there is no unified way to access the mined information through large-scale applications. Typically, different data models are adopted, hindering a simplified access mechanism and integration with external knowledge bases. This fragmentation is not desirable, and text-mining research should encourage the use of modern data exchange standards, allowing researchers to leverage a common layer of interoperability.

Furthermore, the research community has witnessed an impressive growth of biological and biomedical data, collected from multiple research experiments and generated from

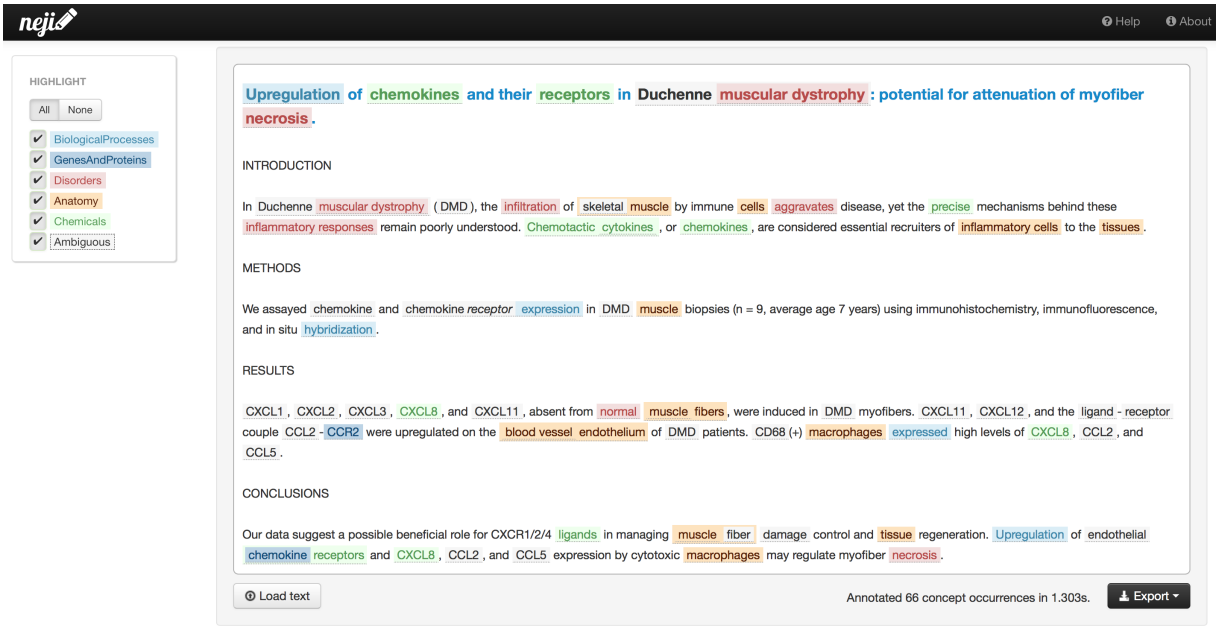


Figure 1.2: Sample extraction of textual information using a computerized web tool (Neji [8]).

daily clinical practice. A significant contribution to these data comes from the increasing availability of sophisticated laboratory equipment. For instance, the use of Next-Generation Sequencing (NGS) technologies, mapping human genetic sequences to digital data, has grown steadily, producing enormous amounts of data [9]. Therefore, it is crucial to make sense of these data to grasp the vast potential for custom drug development and personalized medicine. Their understanding lies in the expertise in several life science fields and connects researchers with distinct requirements and expectations, from gene curators to pharmaceutical researchers and medical clinicians. For those reasons, bioinformatics software solutions must be focused not only on analysis of the raw data, but also on their representation, integration and interoperability, for further use.

In life sciences, most of the available data are fragmented, disregarding good practices for integration. In particular, the integration of heterogeneous data types is currently a very active field of research, where hybrid approaches try to improve data usage and scientific discovery [10]. In that way, data integration remains an open challenge, in which complexity increases with the heterogeneity of data sources. Assorted data sources are difficult to link, connect and share, making it more arduous to interconnect distinct bio-related domains. For instance, connecting distributed information is particularly important in the rare diseases domain, where data needs to be combined from several labs to generate

substantial statistical conclusions. This raises the demand for novel and computer-aided research methods to analyze and connect collected knowledge, reducing human efforts to interpret and understand the information gathered.

In recent years, Semantic Web (SW) [11] has been identified as a common framework to create accessible and shareable information across application and database boundaries. Its adoption by the life science community allows better standards and technologies to be delivered, helping to solve common problems such as data heterogeneity, format diversity and repository distribution. It aims to locate data anywhere on the web, representing a vision in which computers, as well as people, can find, read, understand and use data over the World Wide Web to accomplish useful goals [12]. When independent systems share this type of representation, interoperability and effective data integration across knowledge domains are achieved.

Several semantically-powered databases and services have appeared during the last decade, trying to bring the advantages of SW to the life science community [12], and making the interconnection and exploration of several biological databases possible. One of the earliest projects was the Bio2RDF [13], which collected and converted a variety of biological data, e.g. genes, proteins and pathways, into an accessible triplestore. Recently, the EMBL-EBI RDF Platform [14] was launched, aggregating several biological triplestores including UniProt, ChEMBL, and Reactome. Combined, these triplestores supply a vast amount of semantically-structured information, in which federated inquired mechanisms can easily be applied [15, 16].

In this way, the SW paradigm involves a broad set of modern technologies that are a perfect fit for life sciences' innate connectedness, being able to tackle traditional data issues such as heterogeneity, distribution and interoperability and providing an interconnected network of knowledge.

## 1.2 Research goal

The main objective of this thesis is to **investigate computational methods that facilitate the semantic integration and reuse of biomedical resources**. It aims to advance interoperable solutions, supported by SW concepts and features, with special focus on the development of personalized tools for research on rare diseases. To attain this global purpose, several goals were carefully defined:

- Architect software solutions to integrate and gather knowledge from several

biomedical data sources.

- Research enhanced methodologies for data interoperability, reusability and access.
- Develop improved methods for information distribution and exchange.
- Contribute to international research projects by applying the developed solutions in a specific area, i.e. rare disease research.
- Collaborate with international research groups to promote knowledge sharing and partnerships.

## 1.3 Methodology

To achieve the main objective of this thesis, five separate tasks were defined:

1. **Investigate current challenges associated with biomedical resources' diversification.** In recent years, we have witnessed an explosion of diverse biomedical data sources, resulting largely from the demands of life science research. The vast majority of these data are freely available via diverse bioinformatics platforms in several formats, including biological databases, technical reports and obviously, in the scientific literature information systems. The integration of that heterogeneous information into an interoperable, reusable and shareable infrastructure remains an open challenge, in which complexity increases with the heterogeneity of data sources. Indeed, novel interoperable systems and services are needed to mitigate this problem and to enable innovative knowledge discovery methods.

2. **Enhance current software solutions for data migration.** Biomedical research requires technical infrastructures to deal with assorted data sources. By adopting SW standards and concepts, we are limitless in exploring these data and shaping associations with external resources, avoiding traditional interoperability problems. This results in a flexible transition from traditional systems to an information system sustained by a fully semantic software stack.

3. **Efficient methodologies for information distribution and exploration.** Novel and machine-based strategies are needed to explore the evident value of biomedical

research data interconnection and to enable proper data attribution mechanisms. The development of flexible and effortless data sharing solutions are current requests from the research community, to allow information access and further exploration.

4. **Research flexible and optimized software solutions.** Given the increasing amount of data being published, and the increasing standardization efforts in knowledge representation and exchange, it is vital to develop software solutions that are compliant with SW features and services. Additionally, these solutions require flexible and optimized deployments for the creation of future interoperable bioinformatics platforms.

5. **Validate research methods in a biomedical related domain.** Evaluation of the proposed methods requires the implementation and validation of uses cases in specialized fields. This thesis is specifically focused on Neurodegenerative (NDD) and Neuromuscular (NMD) rare diseases, in which the identified technological weaknesses are underestimated and the connection of distributed resources is vital for medical and research discoveries.

## 1.4 Contributions

During this research, several computational solutions were investigated to fulfill the main research goal, applying new techniques and following innovative approaches to solve or minimize existent concerns in the biomedical community.

The first contribution resulted in the implementation of a semantic layer that allows connecting distributed and heterogeneous rare disease patient registries, giving the opportunity to answer challenging questions across disperse labs [15, 17]. The interconnection between those registries using SW technologies allows queries through multiple instances according to the researcher's needs. Furthermore, the developed **Linked Registries** web-based platform [18] creates a holistic view over a set of anonymized registries, supporting semantic data representation, integrated access and querying.

The second contribution is focused on the enhancement of **COEUS** [19], a SW framework for biomedical data integration [20]. The tool is targeted to support the integration of heterogeneous life science data, providing an intelligent resource combination mechanism. Through advanced developed algorithms, the information can be ported to the semantic level using existing ontologies and models, promoting distributed information

access [21–27]. The conceived web platform performs automated integration of biomedical resources enabling adequate data sharing mechanisms and an efficient attribution process for knowledge exchange.

The third contribution of this thesis is related to the inexistence of a unified strategy to integrate information extraction results into a reusable, shareable and searchable structure [28, 29]. As an outcome of this research, a novel and modular architecture for textual information integration using SW features and services is proposed [30]. Supported by the **Ann2RDF** solution [31], it allows the migration of annotated data into a common model, providing a suitable transition process, in which multiple annotations can be integrated and shared across semantic knowledge bases.

The last contribution is a data migration tool to allow fast and easy transition from traditional information systems to the SW level. Targeted at the biomedical domain, **SCALEUS** [32] is a web-based platform that offers straightforward data migration and SW services. Furthermore, it enables the fast deployment of new semantic-based information systems by including, in a single package, the essential tools needed to contribute to the knowledge federation layer being established across life science research. The solution offers high-performance queries across established networks, delivering a baseline foundation for the creation of shareable and interoperable bioinformatics platforms.

## 1.5 Document structure

The thesis is organized in six more chapters described in Figure 1.3. The main scientific output is also shown for each section.

Chapter 2 provides a state-of-the-art description of the subjects that are most relevant for this work. Strategies for biomedical data integration and distribution are introduced, providing an overview of current methodologies and associated challenges.

Chapter 3 addresses the connection of distributed and heterogeneous rare disease patient registries. The proposed methodology creates a holistic view over the patient registries, supporting semantic data representation, integrated access and querying.

Chapter 4 describes an automated platform to integrate heterogeneous scientific outcomes following adequate guidelines. This results in seamless integration to make data accessible and citable at the same time, without extra scripting methodologies.

Chapter 5 presents a modular architecture to support the integration of text-mined information from independent systems. The presented architecture provides a seamless

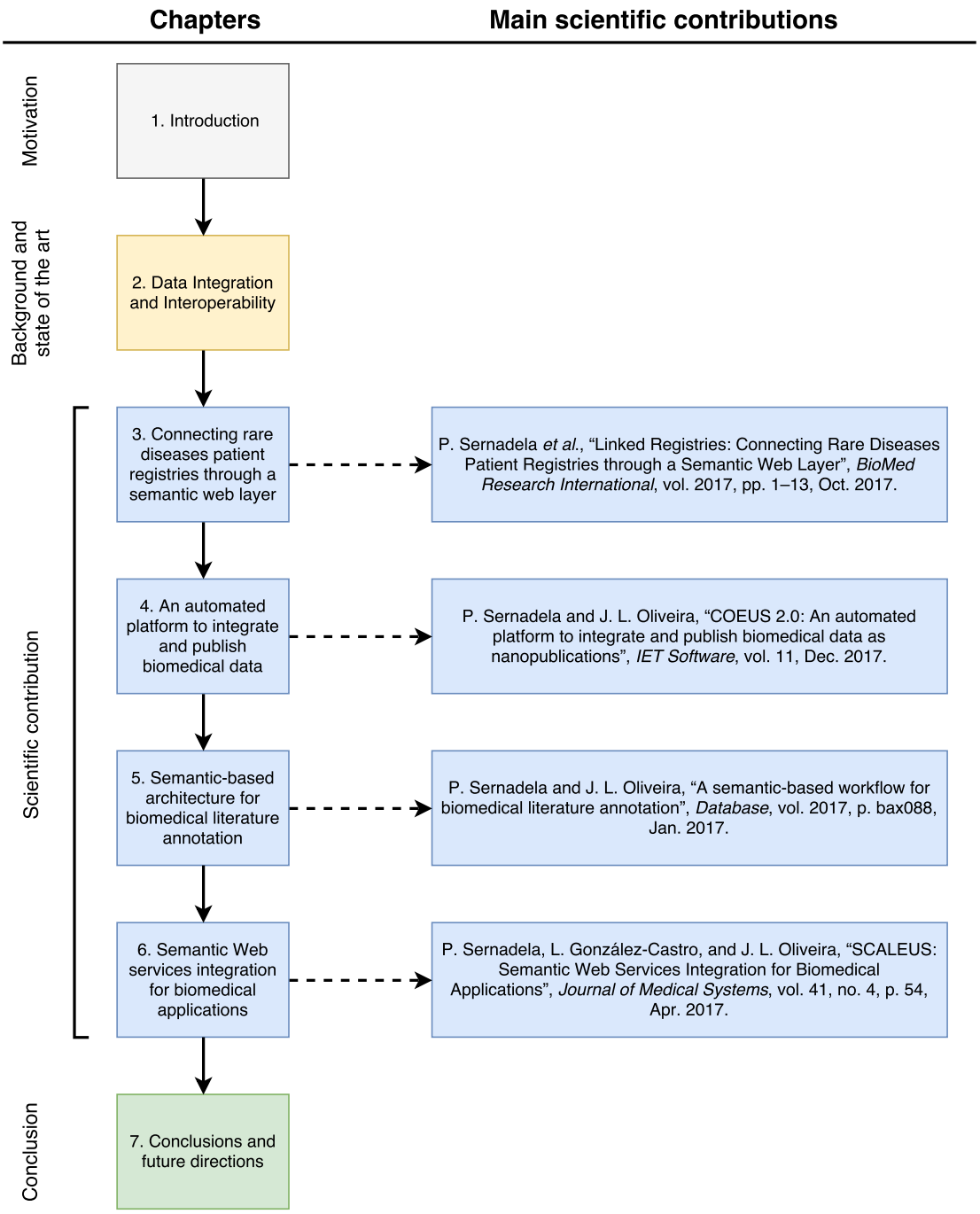


Figure 1.3: Thesis structure, highlighting the main scientific contributions.

transition from unstructured information to the SW level, enabling the full exploitation of curated knowledge according to modern standards.

Chapter 6 presents an open-source platform to facilitate the creation of new semantically enhanced information systems. This web-based system provides rapid data migration methods to foster the adoption of SW technologies.

Finally, chapter 7 presents the final remarks of this thesis, highlighting some directions for future work.

# Chapter 2

# Data integration and interoperability

In recent years, we have witnessed an explosion of biomedical resources resulting largely from the demands of life science research. The vast majority of these resources are freely available via diverse bioinformatics platforms, including relational databases, search engines and scientific literature repositories. The conventional method of individually accessing these resources has achieved great results in the past, but it is unfeasible to support the new paradigm of interoperable open science. Assets need to be integrated and shared among different, scattered sources, reducing the overwhelmingly heterogeneous landscape in the current life sciences ecosystem. This creates novel opportunities to develop methods and technologies to fully extract connected knowledge and fosters the establishment of a linked network of biomedical resources.

## 2.1 Biomedical resources

Current biomedical research benifits from a great availability of biomedical resources. Consequently, the overwhelming growth of bio-related data sources results in more databases, applications, platforms and services. For instance, the Nucleic Acids Research (NAR) Journal yearly tracks the most relevant biosciences database metrics. Examining the Figure 2.1, we can notice the continuous growth of databases in the last 5 years. Moreover, this progression shows not only the development of new databases but also the publication of peer-reviewed articles to describe them in detail. According to a NAR journal report of 2016 [33], there are a total of 1685 different databases that are publicly accessible online. This estimate of publicly accessible databases can be considered conservative. In fact, there are many more online services without complementary

publication in peer-reviewed journals or being developed by commercial companies, making them underrepresented in the scientific community [34].

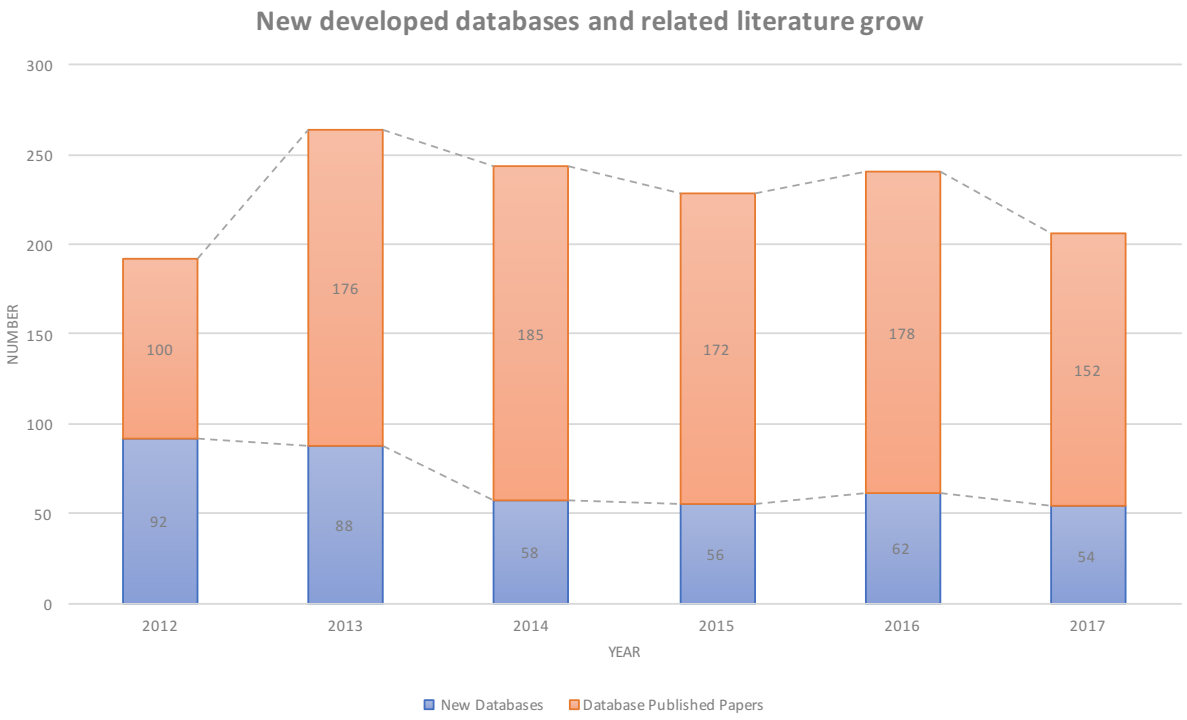


Figure 2.1: Growth of new developed databases and their related publications according to NAR Journal yearly metrics.

Biological databases aggregate vast amounts of omics data, serving as vital resources and becoming increasingly indispensable for scientists, from wet-lab biologists to in silico bioinformatics. They are developed for diverse purposes, covering various types of data and curated at different levels with miscellaneous methodologies. Table 2.1 includes a summarized collection of human-related databases including gene, disease, pathways, drugs and protein databases widely used and currently accessible via the Web. Curation processes are an important part of these databases, involving standardization procedures, quality controls and enhanced data consistency.

Table 2.1: Publicly available biomedical databases.

| **Database** | **Content** |
|---|---|
| Comparative Toxicogenomics Database (CTD) [35] | chemicals, genes, diseases, pathways |
| DrugBank [36] | drugs |
| Entrez Gene [37] | genes |
| Expert Protein Analysis System (ExPASy) [38] | proteins |
| GenBank [39] | nucleotide sequences |
| HUGO Gene Nomenclature Committee (HGNC) [40] | genes |
| Human Metabolome Database (HMDB) [41] | small molecule metabolites |
| Kyoto Encyclopedia of Genes and Genomes (KEGG) [42] | genes, genomes |
| Medical Dictionary for Regulatory Activities (MedDRA) [43] | medical terminology |
| Medical Subject Headings (MeSH) [44] | medical terminology |
| NCBI BioSystems [45] | biological systems |
| Online Mendelian Inheritance in Man (OMIM) [46] | genes, genetic disorders |
| Pharmacogenomics Knowledge Base (PharmGKB) [47] | genes, diseases, drugs |
| Protein Data Bank (PDB) [48] | proteins |
| RxNorm [49] | drugs |
| Systematized Nomenclature of Medicine (SNOMED) [50] | clinical health terminology |
| Toxicology Data Network (TOXNET) [51] | chemicals, environment, toxicology |
| Universal Protein Resource (UniProt) [52] | proteins |

The availability of a vast number of biological databases and related assets poses a major challenge for connecting data sources. Typically, they are physically distributed and heterogeneous in data type and format, requiring specific applications to ease data exchange and sharing. Hence, resource diversification and distribution issues have been key elements slowing down the exploration of biomedical data. Additionally, systems are continuously built disregarding interoperable interfaces, expanding bioinformatics science to a more and more fragmented landscape. This scenario is not desirable, and future software development should be carefully rethought to allow an easy interconnection of bio-related resources.

### 2.1.1 Data diversity

Nowadays, scientists access a wide variety of biomedical resources. Diverse databases, services and applications are accessed through the internet, supporting scientific methods and fomenting life science innovation. To simplify these tasks, data needs to be assembled and processed.

Hence, connecting data sources plays a central role in scientific discovery. In most cases, the task of accessing or integrating multiple data sources is difficult to achieve due to a variety of reasons, e.g. databases do not follow a single model or notation, biological concepts are represented in distinct models and with different identifiers, several types of data formats and structures, among others.

This diversity can be split in 4 main levels. Table 2.2, summarizes those levels' complexity, organizing resource dependencies from data storage heterogeneity to the various access methods.

Table 2.2: Biomedical resources diversity, from data storage heterogeneity to the various access methods.

| **Storage** | **Formats** | **Models** | **Access** |
|---|---|---|---|
| Relational Database<br>Object-oriented Database<br>Textual File<br>Binary File<br>... | HTML<br>CSV<br>XML<br>TXT<br>Excel<br>JSON<br>... | Structure<br>Ontology<br>Semantics<br>... | Local<br>Access<br>Remote APIs<br>Web Service<br>FTP Server<br>Web Pages<br>... |

### Data Storage

Data is stored in many different types of repositories, e.g. in a relational database, NoSQL database, binary file, or usually, in the textual form. These dissimilar repositories have different access boundaries, making resource integration complex and arduous to perform. For instance, accessing data in a relational database such as MySQL is totally different from accessing a remote flat file from a File Transfer Protocol (FTP) server: different interfaces need to access different methods to get the desired content.

### Data Formats

Even if data storage is performed in the same physical format, such as flat file, these files can vary in data formats. Accessing procedures are distinct for the examination of a Comma-Separated Values (CSV) and Extensible Markup Language (XML) file, for instance. This generates great heterogeneity in read/write operations, making it necessary to assimilate different types of syntax to allow data integration.

### Data Models

At the data model level, there can also be structure or schema heterogeneity. For instance, exploring the XML standard, several concerns can be found with available format structures: although there are normalization processes for read and write procedures, different structures can be found, diverging from application to application. These issues can be resolved by adopting some information mapping techniques, usually requiring extra complex and demanding tasks.

### Data Access

Finally, data access methods are also challenging to integrate if different Web services or protocols are used. For instance, using different protocols such as Hypertext Transfer Protocol (HTTP) and FTP requires different methods for remote data access, making it necessary to adapt data requests. Generally, resource heterogeneity requires the development or adaption of integrative software systems to foment interoperability demands.

### 2.1.2 The rare disease landscape

A rare disease is a particular health condition affecting almost 1 in 2000 people. [53]. According to the Orphanet inventory (`www.orpha.net`), there are approximately 6000 to 8000 rare diseases, of which about 80% have a genetic origin. Complex health implications behind rare diseases are seldom considered in medical or social care. Due to the rarity of each individual disease and their often complex nature, this group is underrepresented in research and treatment developments. At the patient level, the diagnosis of a rare disease generally means more difficulty in finding both clinical and psychological support [54]. The existence of a small number of cases for each disease creates additional barriers in the translational research pathway, as it is difficult to identify and coordinate a substantial cohort [55, 56]. Nevertheless, altogether rare disease patients comprise an estimated 6 to 8% of the EU population [57].

During the last decade, several small disease-specific databases have been developed, related, for instance, to neurological disorders or muscular problems [58]. While they provide high quality information and resources, their coverage is small and typically with a regional or national scope. To achieve greater statistical evidence, we need extensive cohorts of patients with similar features, from a worldwide population. Hence, discovering rare disease-causing genes and mutations can have an impact on all medical treatment stages, from clinical diagnostics to insights gained into biological mechanisms and common diseases [59]. In addition to long-term patient care improvements, understanding gene-disease associations is a fundamental goal for bioinformatics research, especially in rare diseases where genotype-phenotype connections are typically limited to one or a few genes [60, 61]. Moreover, it is in these particular conditions that the strongest relations between genotypes and phenotypes are identified. Hence, to fully understand the underlying causes of diseases, we need to connect knowledge that is widespread throughout miscellaneous registries.

#### Patient registries

The collection and maintenance of patient registers has assumed a key role in the identification of new treatments and in the improvement of care. In particular, personal genetic records are of growing interest. These data are increasingly important for diagnosis, resolution and therapeutic treatment of rare diseases. Hence, databases with information about the human genome, such as the Human Gene Mutation Database (HGMD) [62] or the 1000 Genomes Project [63], are increasingly relevant. Moreover, it is important to reuse

these data in novel biomedical software to enable their use in daily medical workflows.

The value of individual data increases when it is aggregated and presented in a unified way, both for humans and computers [64]. Orphanet provides a public portal, for professionals and patients, with the most up-to-date information about rare diseases and orphan drugs [65]. It also displays information on specialized consultations, diagnostics, research projects, clinical trials and support groups. Diseasecard [66] is another platform that aggregates genotype-to-phenotype information regarding rare diseases, pointing to key elements for both the education and the biomedical research field. While these systems do not provide repositories for patient-level data, they are useful resources for sharing and disseminating existing knowledge and expertise.

Besides the important role of these specialized repositories, the integration of knowledge that can be extracted from distinct Electronic Health Record (EHR) is also a major challenge to support personalized medicine. Data from gene sequences, mutations, proteomics and drug interactions (the genotype) can now be combined with data from EHRs, medical imaging, and disease-specific information stored in patient registries (the clinical phenotype). Hence, it is crucial to start exploring patient-level data from rare disease registries, which often include personal data, diagnosis, clinical features, phenotypes, genotypes, treatments and clinical follow up.

According to Orphanet, there are over 600 rare disease registries just in Europe, with different aims and objectives, with access to different resources and collecting different datasets. Registries have traditionally been developed to accelerate the translational research pathway helping to move therapies from bench to bedside as quickly as possible. They provide a tool for the feasibility and planning of clinical trials as well as a means to identify and recruit patients for research. However, the purpose and utility of registries have a much broader reach, providing a source of natural history data and a basis for hypothesis generation that can advance research in a given field.

A single researcher or a clinician with an interest in a particular field can set up a registry, or a disease network, or - as is increasingly common - a patient organization. The variety in origin explains the variety of funding schemes (sustainability models) and data collection techniques [67].

The most developed registries (e.g., Cystic Fibrosis [68, 69]) act as detailed studies, with data collected at fixed time points in the clinical setting and stored in bespoke software solutions. However, many registries are online self-report systems, with patients entering data through a web portal.

There are also examples of a combined approach: patients initiate registration, while physicians verify details through the same web portal. This disparity in data collection increases the complexity of a unified system. The data items themselves are not standardized across all rare diseases though a significant amount of effort has been applied in this area. Some consensus has been reached in certain disease areas, such as Duchenne Muscular Dystrophy (DMD), where a federated registry system exists under the umbrella of TREAT-NMD (`www.treat-nmd.eu`) [70, 71].

### Towards harmonization

One step towards harmonisation can come in the form of international medical classifications or languages, such as Unified Medical Language System (UMLS) [72] or SNOMED-CT [73]. More recently, the use of phenotype ontologies, such as Human Phenotype Ontology (HPO) [74], has been proposed for phenotype standardization. Ontologies are a structured representation of knowledge using a standardized controlled vocabulary for data organization, searching, and analysis. The use of ontologies to identify and annotate data in patient registries ensures interoperability and common access, and it also enables cross-cohort comparisons and filtering. Importantly, it allows the development of new bioinformatics tools, covering the automated and systematic matching of clinically similar representations of phenotypes to assist in differential diagnosis, among others.

These patient-centric databases offer unique specialized views over their internal datasets. However, while there are huge amounts of data scattered throughout multiple stakeholders, they are extremely difficult to obtain or access. The main reasons are the lack of semantic compatibility and the evident low motivation of data owners to share and spread data, and thus, individual efforts remain isolated. This is a critical obstacle in rare disease research, where a sole center may collect only a small number of patients with a certain disease. The outcome of this is that, in the end, there is not enough data to generate statistically meaningful conclusions. As such, we cannot discover or infer new knowledge because there is no access to a minimal amount of patient data.

To cope with these challenges we need a platform that offers a unique holistic view promoting the collaboration of multiple entities towards the study of rare diseases and assessment of patients' evolution [75]. According to our study, only one related exchange platform was developed regarding the rare disease domain: The Matchmaker Exchange [76]. This platform provides a systematic approach to create a federation network of genotypes and phenotypes databases through a common Application Programming

Interface (API). This helps in the process of finding common genotype/phenotype pairs in multiple individuals. However, this approach requires depositing the data in the main database or the setup of local instances, always ensuring a set of services and end user agreement. Indeed, generic solutions supporting the creation of independent systems that can be plugged into any existing patient registry without changing it are a better milestone towards semantically interoperable rare disease knowledge.

## 2.2 Semantic web

The first generation of the Web, which started during the 90s, is very different from the Web of 2017. In the beginning, it was mostly about publishing static Hypertext Markup Language (HTML) pages into a server, using rudimentary edition mechanisms.

The second generation of the Web was driven by a more dynamic and interactive content: people search on the Web, discover answers to complex problems, find friendship and communities, and more [77]. Web 2.0 influenced millions of people, creating new research and business opportunities and initiating the era of social networks and online advertising.

The third generation, the Semantic Web (SW) [11], is all about improvements in the connectedness of Web 2.0, aiming to make data located anywhere on the web accessible and understandable, both to people and to machines. In this way, it stands for a vision in which computers, as well as people, can find, read, understand and use data over the World Wide Web (WWW) to accomplish useful goals for users [12]. Overall, it is based on a new technology that helps to reuse and repurpose data on the Web in new ways, following a set of key characteristics:

- *Ubiquitous networking* - Data needs to be connected irrespective of its physical location. Networks must remain open. Open data, open services, open APIs, open protocols, and formats are the vision of the Web 3.0.
- *Adaptive information* - Data and resources are becoming increasingly more connected and more dynamic, which implies reassembling on demand.
- *Adaptive services* - The Semantic Web movement requires the publication and consumption of data *as a service* [78] hosted via Web protocols.
- *Federated Data* - Data needs to be stored and retrieved from different locations during a single query.

- *Web Intelligence* - Semantic networks make available description logic algorithms, such as reasoning, to the Web. The logical formalism in the SW allows the extraction of useful meaning from the data, automating the way people interact with it.

While SW is just a data recombination paradigm over the Web, a lot of work still needs to be done to take full advantage of it and to allow a world-wide linked Web.

### 2.2.1 Linked Data

The web has advanced from an overall information space of connected pages to one where both pages and content are linked, enabling a truly distributed knowledge network. In this evolution process, a set of principles and standards were defined to enable structured data publishing on the web. In 2006, Tim Berners-Lee [79] outlined four of these best practices:

- Use Uniform Resource Identifier (URI)s as names for things;
- Use HTTP URIs so that people and machines can search for those names;
- When someone looks up a URI, it provides useful information;
- Include links to other URIs so that they can discover more things.

Despite being somewhat vague, following these principles allows an easily navigable distributed data graph. Thus, the term Linked Data is simply about creating links between diverse resources [80]. One relevant example is the Linked Open Data (LOD) [81], an initiative which aims to deliver Linked Data under an open license. In recent years, the project has grown considerably forming a global data space containing billions of assertions, the Web of Data. DBpedia [82], one such initiative, is a central hub that provides a huge knowledge base with data extracted from Wikipedia, and currently includes over 400 million facts describing 3.7 million things.

By publishing on the web according to these best practices, data became incorporated into a global space allowing them to be discovered and used by several applications.

### 2.2.2 General concepts

The fundamental unit of the Semantic Web (SW) knowledge is a *statement*, or *triple*, a single piece of metadata formed through the union of three elementary components: a

*subject*, a *predicate* and an *object*. Examples of *statements* are: *"P05067 - is a - Protein"*, *"Pedro - lives in - Aveiro"* and *"Bragança - is located in - Portugal"*. The *subject* is the element where we will apply something new (e.g. *P05067, Pedro, Bragança*), the *predicate* is the meaning of the relationship established (e.g. *is a, lives in, is located in*), and the *object* is the explicit target (e.g. *Protein, Aveiro, Portugal*). Based on this relation, *statements* can be linked together, forming successive chains of links. The possibility to create such numerous relationships and interactions is the SW's main added value. Connecting millions of *statements* together can form a rich knowledge base, even if the information remains distributed. These relationships are commonly understood as a graph, as shown Figure 2.2 with a small set of statements regarding the *P05067* protein.

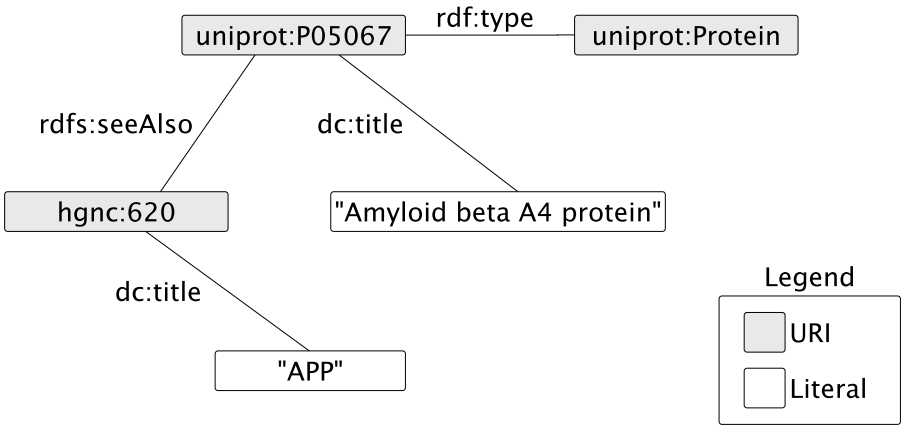


Figure 2.2: A set of statements related with the UniProt *P05067* protein.

Each resource must be uniquely identified in a specific *namespace*, an identity space on the Internet, based on the URI definition [83] to identify unique resources across the entire web. For instance, the "Amyloid beta A4" protein can be identified through the combination of UniProt namespace (`http://www.uniprot.org/uniprot/`) with the protein accession number *P05067*, resulting in `http://www.uniprot.org/uniprot/P05067`.

### RDF

The Resource Description Framework (RDF) is a standard model for representing information in the web [84]. For instance, the contact information of a person can be represented using the following RDF sample:

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

  <foaf:Person rdf:about = "http://www.w3.org/People/EM/contact#me">
   <foaf:name>Eric Miller</foaf:name>
   <foaf:phone rdf:resource="tel:+1-(617)-258-5714"/>
   <foaf:mbox rdf:resource="mailto:em@w3.org"/>
  </foaf:Person>
</rdf:RDF>
```

In this case, the information contact of *Eric Miller*, one of the leaders of the World Wide Web Consortium (W3C) SW initiative, is shown. To define the RDF vocabulary, a Friend Of A Friend (FOAF) ontology (`http://xmlns.com/foaf/spec/`) is used, with it being possible to express metadata about people such as name (e.g. *foaf:name*), phone number (e.g. *foaf:phone*) and email address (e.g. *foaf:mbox*).

According to W3C, the design of RDF was intended to meet the following goals [85]:

- *Simple data model*: the underlying structure of any expression in RDF is a collection of *triples*.
- *Formal semantics and provable inference*: RDF has a formal semantics providing a basis for defining reliable rules of inference in RDF data.
- *Extensible URI-based vocabulary*: the vocabulary is fully extensible, being based on URIs with optional fragment identifiers, i.e. RDF references. URI references are used for naming all kinds of things in RDF.
- *XML-based syntax*: RDF has a recommended XML serialization form.
- *XML schema datatypes*: RDF can use values represented according to XML schema datatypes, thus assisting the exchange of information between RDF and other XML applications. Datatypes are used in the representation of values such as integers, floating point numbers and dates.
- *Anyone can make statements about any resource*: to facilitate operation at the Internet scale, RDF is an open-world framework that allows anyone to make statements about any resource.

Following these goals, there is no limit to exploring data and connecting them with external resources without the traditional interoperability issues. For instance, RDF has features that facilitate data merging even if the underlying schemas differ. This is one of the major advantages of using RDF graphs, facilitating data combination and allowing this to be shared across the web.

### Ontologies

Ontologies define the collection of terms and relations between terms that are adequate for a given topic [86]. These relationships, often designated axioms, establish connections between terms that mimic the real world. An ontology provides the means to classify the "things", to give classification names and labels, and to define the kind of properties and relationships that can be assigned. In practical terms, ontologies are used to assert facts about resources described in RDF and referenced by a URI.

The Web Ontology Language (OWL) [87] is the W3C ontology standard, extending the RDF schema. It is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL offers greater machine interpretability of web content, providing additional vocabulary along with formal semantics. In this way, ontologies are OWL documents, that can be published to the WWW, defining resources and their relationships. For instance, the FOAF ontology already mentioned is used to describe people and social relationships. FOAF is devoted to linking people and information, integrating social networks of human collaboration, friendship and associations. In FOAF descriptions, there are several kinds of things and links, i.e. properties. The types of the things are named classes. FOAF is therefore defined as a dictionary of terms, each of which is either a *rdfs:Class* or a *rdf:Property*. Additionally, other projects can provide other sets of classes and properties, many of which are linked with those defined in FOAF (e.g. *owl:equivalentClass*). As an example, an overview of the FOAF ontology used to describe the *Eric Miller* contact information is shown below:

```
<rdf:RDF
 <!-- RDF namespaces -->
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 xmlns:owl="http://www.w3.org/2002/07/owl#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/"
 ... >
 <!-- FOAF Ontology definition -->
 <owl:Ontology rdf:about="http://xmlns.com/foaf/0.1/"
  dc:title="Friend of a Friend (FOAF) vocabulary"
  dc:description="The Friend of a Friend (FOAF) RDF vocabulary,
   described using W3C RDF Schema and the Web Ontology Language." >
 </owl:Ontology>

 <!-- FOAF classes (types) -->
 <rdfs:Class rdf:about="http://xmlns.com/foaf/0.1/Person"
  rdfs:label="Person" rdfs:comment="A person." >
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class" />
  <owl:equivalentClass rdf:resource="http://schema.org/Person" />
  <rdfs:subClassOf>
   <owl:Class rdf:about="http://xmlns.com/foaf/0.1/Agent"/>
  </rdfs:subClassOf>
  ...
  </rdfs:Class>
  ...
 <!-- FOAF properties -->
 <rdf:Property rdf:about="http://xmlns.com/foaf/0.1/name"
  rdfs:label="name"
  rdfs:comment="A name for some thing.">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#
   DatatypeProperty"/>
  <rdfs:domain rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#
   Literal"/>
  <rdfs:isDefinedBy rdf:resource="http://xmlns.com/foaf/0.1/"/>
  <rdfs:subPropertyOf rdf:resource="http://www.w3.org/2000/01/
   rdf-schema#label"/>
 </rdf:Property>
 ...
</rdf:RDF>
```

Regarding the biomedical domain, there are also ontologies for terminology description. To exemplify, Table 2.3 includes a summarized collection of biomedical ontologies currently available on the web.

Table 2.3: Some publicly available biomedical ontologies.

| **Ontology** | **Content** |
|---|---|
| Cell Ontology (CL) [88] | cells |
| Chemical Entities of Biological Interest (ChEBI) [89] | chemicals |
| Disease Ontology (DO) [90] | diseases |
| Drug-Drug Interactions Ontology (DINTO) [91] | drug-drug interactions |
| Gene Ontology (GO) [92] | genes |
| Protein Ontology (PRO) [93] | proteins |
| Sequence Ontology (SO) [94] | genomic annotations |
| Unified Medical Language System (UMLS) [72] | biomedical terminology |

### 2.2.3 Storage

RDF stores have key distinguishing features compared to the relational databases [95]. Firstly, they are more flexible than relational models, which need to be reorganized if the database schema changes. Second, RDF resources are identified by a unique URI, making it possible to create references between two different RDF graphs, even in different namespaces, therefore enabling data linkage. Third, the relational model does not have the notion of hierarchy, which makes it difficult to apply Structured Query Language (SQL) queries for reasoning purposes. In opposition, this type of query is natively supported in RDF Schema (RDFS) and OWL. As a result, RDF repositories offer easier data integration of diverse sources as well as more analytical power.

### Triplestore

A triplestore is a type of graph database that allows storage and retrieval of facts through semantic queries. Being a graph database, a triplestore stores data as a network of objects with materialized links between them, i.e. *triples*. This makes triplestores the preferred choice for managing highly interconnected data. They are also capable of handling powerful semantic queries and using inference to uncover new information. Due to the ability to manage unstructured and structured data in several domains, such as life sciences, these RDF databases are growing very fast [95]. Usually, they differ in characteristics such as scalability, performance, data management methods, reasoning capabilities and licensing, among others.

The first proposed triplestores based their data storage mechanism on top of traditional Relational Database Management System (RDBMS). Initially, this approach allowed faster development with little programming effort. However, the flexibility of the RDF model is poorly suited to traditional relational storage models which, for efficiency reasons, rely on well-defined structural models [96]. One of the difficulties in implementing triplestores on SQL databases is that, while it is possible to store the triples, it is very difficult to implement efficient queries on a RDF-based graph over traditional SQL queries. Some of the solutions that follow the relational-based approach are Virtuoso [97], COEUS [98] and Sesame [99].

Therefore, the trend in managing RDF data has moved away from the relational approach to other storage schemas. These new triplestores avoid dependence on rigid SQL schema, and are better suited to the flexible structure of the RDF data. Native triplestores including AllegroGraph [100], BlazeGraph (`https://www.blazegraph.com/`), Fuseki (`http://jena.apache.org`) or Sesame (which has a hybrid approach), are built from the ground up as database engines, allowing exploitation of the RDF data model to store and access RDF data efficiently and showing higher overall performance. In this way, several RDF stores are currently available and each of them may be more suitable in specific cases, depending on the requirements of the scenario [101]. For instance, Table 2.4 presents a brief comparison of several systems currently available for the creation and management of RDF repositories. This overview was assessed by informally measuring the installation and configuration effort (i.e. setup), type of API, storage method, inference support and type of license.

Table 2.4: Synopsis of some available SW applications for triplestore creation and management.

| APP | Setup | API | Storage | Inference | License |
|---|---|---|---|---|---|
| Virtuoso | Complex | REST/SOAP | RDBMS | RDFS | Open Source |
| Sesame | Complex | REST | Native / RDBMS | RDFS + Rules | Open Source |
| COEUS | Complex | RESTstyle | RDBMS | No | Open Source |
| Fuseki | Complex | REST | Native | No | Open Source |
| Allegrograph | Complex | REST | Native | RDFS | Commercial |
| Blazegraph | Easy | REST | Native | RDFS | Open Source |

### SPARQL

SPARQL Protocol and RDF Query Language (SPARQL) [102] is the standard query language for RDF. It enables users to query information from databases or any data source mapped to RDF. It is similar to SQL query language, allowing the user to retrieve and modify data. SPARQL syntax is also very similar to SQL and enables five distinct types: *SELECT*, *CONSTRUCT*, *ASK*, *UPDATE* and *DESCRIBE*.

*SELECT* statements are similar to SQL selections where we bind variables in our query to the results we expect to obtain from the database. The basic structure of a *SELECT* query comprises the *prefix* declarations, for abbreviating URIs, the *result clause*, to identify what information to return from the query, and the *query pattern*, to specify what to query in the underlying database. For instance, the following query shows an example of how to retrieve the name(s) and email(s) of a RDF database using SPARQL:

```
# prefix declarations
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>
# result clause
SELECT ?name ?mbox
# query pattern
WHERE {
    ?person foaf:name ?name .
    ?person foaf:mbox ?mbox
}
```

Additionally, *CONSTRUCT* queries can be performed instead of *SELECT* statements, providing an alternative result clause, i.e. instead of returning a table of result values,

it returns a RDF graph. *ASK* queries evaluate the existence of a particular resource or relationship, returning a boolean value. Finally, *UPDATE* queries allow updating data in a knowledge base graph, while *DESCRIBE* queries return all known relationships for the given resource. SPARQL queries are directed to a *SPARQL endpoint*, a service that accepts queries and returns the results in one or more machine-processable formats. Besides being the SW query language, SPARQL is also the protocol for setting up HTTP connections from clients to endpoints. SPARQL *endpoint* becomes the main preference to access data because it is a flexible way to interact with the Web of Data, by formulating queries like SQL in a traditional database. In contrast to SQL, SPARQL queries are not constrained to work within one database. Based on data source location, the infrastructure for querying Linked Data can be divided into two main categories: central and distributed repositories. The central repository has the same characteristic of data warehousing, where the data are collected in advance in a single repository before regular query processing. In contrast to query distributed repositories, federations-based solutions must be adopted [103]. With these federation systems, the data is discovered by following HTTP URIs of distributed endpoints, each distinct repository providing a wide and heterogeneous query engine that supports the principles of Linked Data. This type of federation strategy has been the topic of recent research in the SW research community [16]. For instance, we can use the SPARQL Federated Query specification (`https://www.w3.org/TR/sparql11-federated-query/`) to execute distributed queries over different SPARQL endpoints. This is performed by using the *SERVICE* keyword to instruct a federated query processor to invoke a portion of a SPARQL query against a remote SPARQL endpoint. The next example shows how to query a remote SPARQL endpoint (to find the names of the people we know) and join the returned data with the data from a local RDF database:

```
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>
SELECT ?name
FROM <http://example.org/myfoaf.rdf>
WHERE
{
  <http://example.org/myfoaf/I> foaf:knows ?person .
  SERVICE <http://people.example.org/sparql> {
    ?person foaf:name ?name .
  }
}
```

The growing number of SPARQL query services offer data consumers an opportunity to merge data distributed across the Web. These services allow effective use of data in a universal and machine-understandable way. For that reason, users must cooperate and deploy such interoperable services avoiding future deep modifications to satisfy data integration demands. Nevertheless, this cooperation can only bring successful results if well-described and controlled content is provided.

### 2.2.4 Technology adoption

Life science research is continuously pushing forward novel strategies capable of significantly improving the research workflow. To cope with current demands, bioinformatics researchers are adopting emerging SW [11] technologies to achieve better solutions to represent and analyze biological and medical processes. The complex relationships behind such processes are easily mapped onto semantic graphs, enabling greater understanding of collected knowledge. Latest advances in the area cover the research and development of new algorithms to further improve how we collect data, transform data into meaningful knowledge assertions, and publish connected knowledge.

The majority of life science data are scattered through closed independent systems, disregarding any good practice for integration and interoperability features [104]. Moreover, the overwhelming scale and intrinsic complexity of the data generate information overload, requiring additional efforts to gather insights from the available knowledge [105, 106]. Additionally, the role of bioinformatics researchers is affected due to the use of heterogeneous tools to attack each specific problem. The use of different systems and applications creates communication issues, resulting in a significant ecosystem fragmentation.

With adoption of the SW paradigm, new standards and technologies allow the solution of common problems, from information heterogeneity to knowledge distribution [107]. From a technological standpoint, the SW can be seen as an "intelligent" data network, enabling meaningful relationships amongst data. As such, the SW emerges as a next-generation software development paradigm able to combine life science characteristics with integration and interoperability demands, providing improved computational features to exchange and accurately interpret knowledge [108]. As with the majority of new technologies, updating or migrating existing systems to a new working environment are cumbersome tasks. Likewise, moving systems from relational repositories, or even from flat files, to semantic infrastructures has been the subject of extensive research [109–111]. Evolving

such systems to the SW ecosystem is a trend that seems likely to continue in the coming years. Overall, major emphasis has been given to the development of translation languages and algorithms, enabling the mapping from data connections to the SW graph.

Combining biomedical data with SW features allows the extension of existing relational connections, enriching their meaning and expressiveness. Regarding this subject, two approaches are common: some mappings are dedicated to forming new triple sets from existing relational databases, whereas other languages enable publishing semantic views over relational data. These languages are complemented with translation applications, using the newly mapped model to provide a semantic data version. Triplify [112] and D2R server [113] are examples that allow semantic views over existing relational data. Despite these advances in migration technology, the resulting systems are just a semantic version of pre-existing relational data. For this reason, other systems explore different challenges in the integration and transition process. For instance, Bio2RDF [13] was one of the first to successfully integrate heterogeneous resources from the most relevant life science databases, from genes to proteins up to pathways and publications, and deliver semantic services to access this information. A more recent example is COEUS [20], an open source framework whose target is to streamline the development cycle of SW applications. The framework provides a single package including advanced data integration and triplification tools, base ontologies, a web-oriented engine and a flexible exploration API. Resources can be integrated from heterogeneous sources, including CSV and XML files or SQL query results, and mapped directly to one or more ontologies. With the same decision-support goals as traditional warehouse systems, these applications adopt advanced Extract-Transform-and-Load (ETL) techniques to triplify existing data into a semantic format, storing them in triplestores.

## 2.3 Information extraction

The continuous growth of scientific literature repositories demands the exploration of automated information extraction tools to access relevant information contained in millions of textual documents and to support translational research [114]. In the biomedical domain, progress has been outstanding [115], producing reliable text-mining tools and innovative text-processing algorithms. The combination of these techniques has been increasingly applied to assist bio-curators, allowing the extraction of biomedical concepts such as genes, proteins, chemical compounds or diseases, and thus reducing curation times and

cost [5]. The manual curation of these data is a demanding task, and the latest strategies use computerized text-mining solutions to aid in the analysis, extraction and storage of relevant concepts and their respective attributes and relationships. Recently, interactive solutions have attracted more attention due to the added benefits of including automatically extracted information in the manual curation processes. With these solutions, the curation time is improved and possible mistakes from computational information extraction results are minimized. *Brat* [116], *MyMiner* [117], *Argo* [118] and *Egas* [119] are examples of interactive solutions, aiming to simplify the annotation process. For instance, *Egas* splits the document text into highlighted sentences to simplify visualization and promotes focused text analysis and improved information extraction processes (Figure 2.3). Accordingly, it also supports in-line information annotation of concepts and relations.

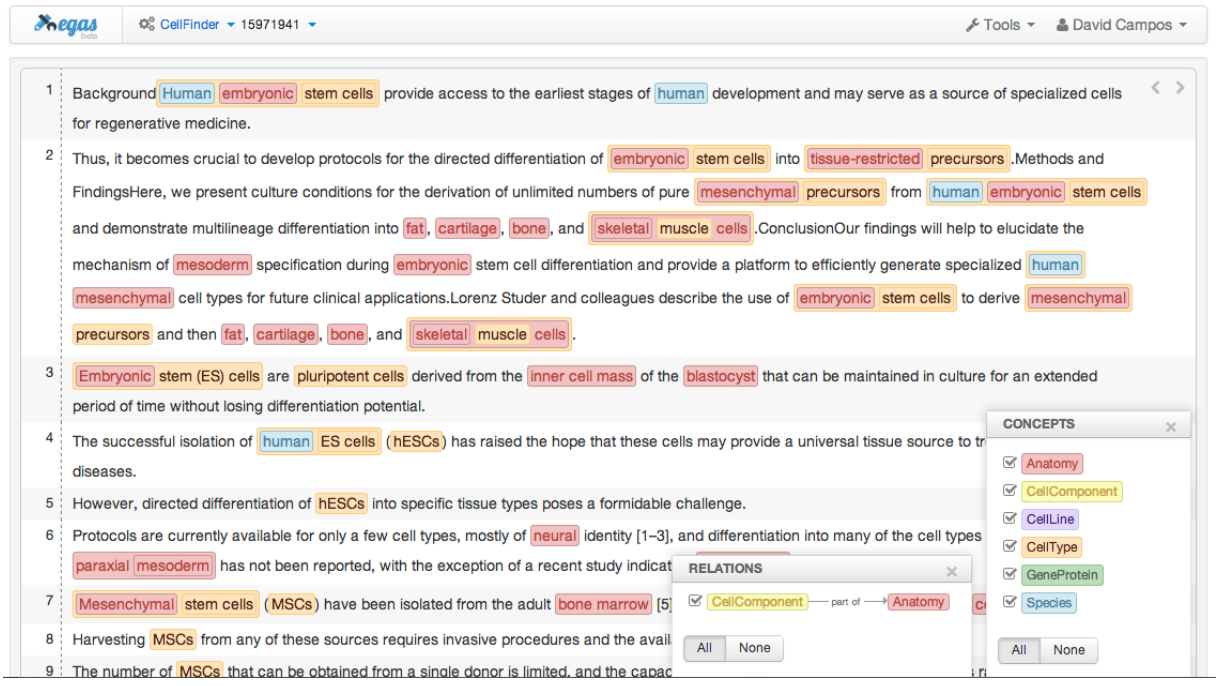


Figure 2.3: *Egas* tool [119] annotating concepts and relations in a sample text document.

Indeed, biomedical information extraction aims to extract information from text documents, such as abstracts, articles, documents and reports. To do so, state-of-the-art solutions usually follow a combination of pre-defined and sequential processes, illustrated in Figure 2.4.

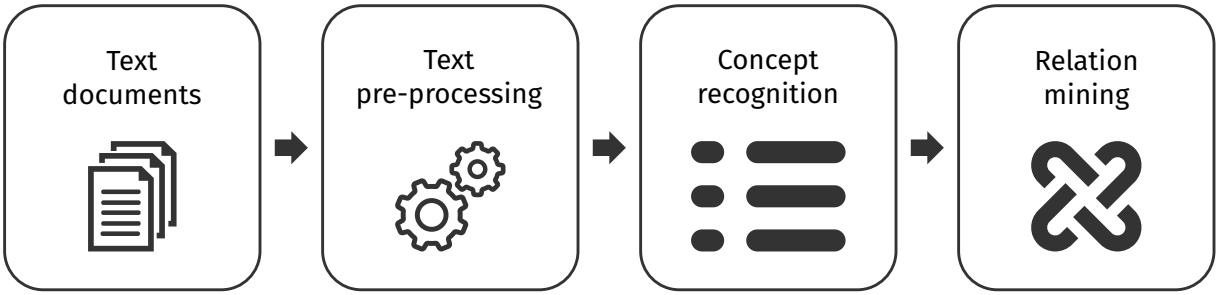


Figure 2.4: Main processing steps currently applied in biomedical information extraction.

### 2.3.1 Corpora and evaluation

To evaluate biomedical information extract systems, it is fundamental to compare them with existing solutions. Usually, this is performed by comparing the automatic annotations output with the annotated corpus provided by expert curators. An annotated corpus, by definition, is a set of documents that usually contains annotations of specific domains. Table 2.5 shows some biomedical annotated corpora that can be used to evaluate biomedical IE solutions.

Table 2.5: Some publicly available biomedical corpora.

| **Corpus** | **Purpose** |
|---|---|
| CRAFT corpus [120] | concept recognition |
| GENIA corpus [121] | concept recognition |
| Gene Regulation Event Corpus (GREC) [122] | event extraction |
| MSH WSD data set [123] | word sense disambiguation |

To measure performance, some specific metrics of the predicted annotations should be calculated. Predicted annotations can be classified as true, if they agree with the correct annotations, or false, if they are not in compliance with the correct annotations. Also, predicted annotations can be classified as positive, if the system provides an annotation, or negative, if the system does not provide any annotation, that is, there is no annotation.

Therefore, predictions can belong to four distinct classes:

- True Positive (TP) = correctly identified;
- False Positive (FP) = incorrectly identified;
- True Negative (TN) = correctly rejected;
- False Negative (FN) = incorrectly rejected.

Several metrics are normally used to evaluate the performance of this classification problem: *precision*, *recall*, *accuracy*, and *F-measure*. These metrics assume values between 0, in the worst case, and 1, in the best case. The *precision* is given by the ratio between the correct predicted annotations, TP, and the amount of predicted annotations, TP+FP (Equation (2.1)).

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

The *recall*, or *sensitivity*, is defined as being the ratio between the correct predicted annotations and total curated annotations (Equation (2.2)).

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

The *accuracy* is defined as the ratio between the true predictions and the total number of predictions (Equation (2.3)).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

Finally, the *F-measure*, or *F-score*, is the harmonic mean of *precision* and *recall* metrics (Equation (2.4)).

$$F\text{-}measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

### 2.3.2 Text pre-processing

Natural Language Processing (NLP) is a field of computer science concerned with the processing of natural language data. NLP began in the 1950s as the intersection of artificial intelligence and linguistics [124] studying problems associated with the automatic generation of text, or speech, and the understanding of human language. Nowadays, NLP techniques can be effectively accomplished by computerized systems, which split

documents into meaningful components, such as sentences and tokens, assign grammatical categories (a process named part-of-speech tagging), and apply linguistic parsing to identify the structure of each sentence. The pipeline can be illustrated by the Figure 2.5, which contains several linguistic tasks.

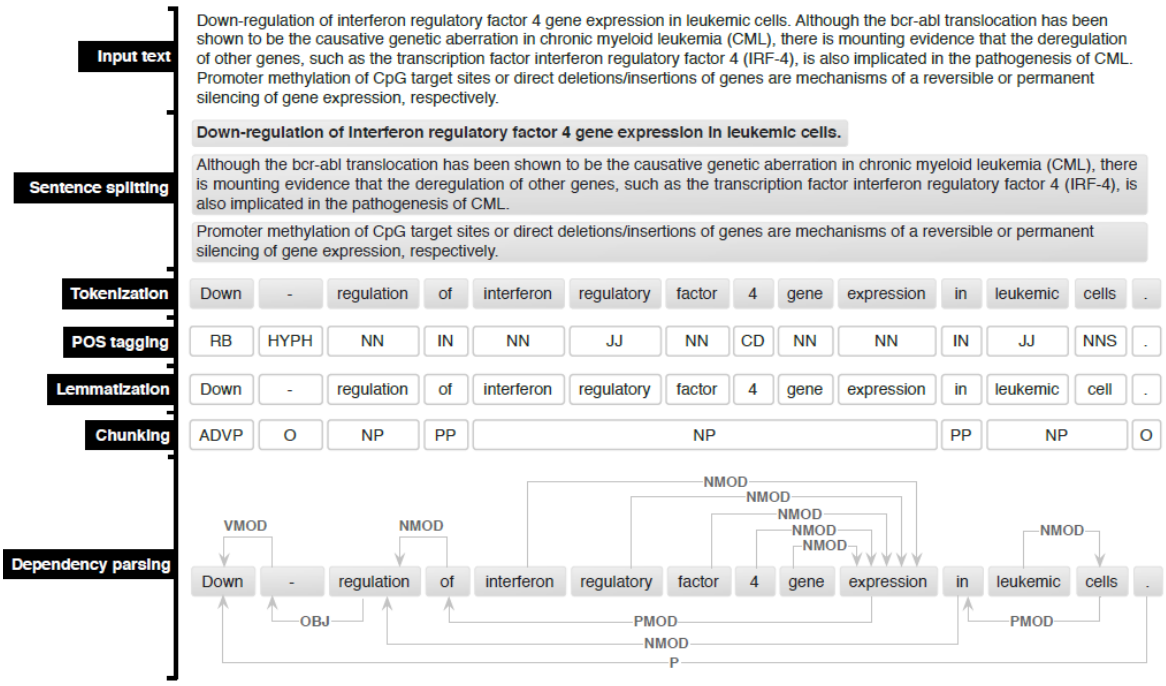


Figure 2.5: Different NLP tasks with respective dependencies considering the sentence "Down-regulation of interferon regulatory factor 4 gene expression in leukemic cells" [125].

**Sentence splitting** goal is to break a given text in the respective sentences. This is acquired using punctuation symbols, such as the period (.), the exclamation mark (!), the question mark (?), and others. The best performing solutions can achieve an accuracy of around 99% using Conditional Random Field (CRF) models [126].

**Tokenization** aims to identify and separate the words from a text, or from the split sentences (if the text was split into sentences firstly). Each separated word, or set of words, is called a token. Barrett and Weber-Jahnke [127] built a biomedical tokenizer that combines regular expressions and ML techniques, achieving accuracies around 92%. A comparison of 13 distinct tokenizers using MEDLINE abstracts was made by He and Kayaalp [128].

A **Stop words** process can be used to remove common words that do not give any

relevant information in the text (e.g., *"and"*, *"so"*, ...).

The **part-of-speech (PoS) tagging** labels each word according to its lexical category (also named word category or word class). Examples of these categories are: nouns, verbs, adjectives and adverbs. Part-of-speech taggers for biomedical texts, can achieve a precision of 97% [129].

The **stemming** process goal is to remove the suffixes from the words (e.g., *"cars"* → *"car"*, *"heroine"* → *"hero"*, *"running"* → *"run"*) and the **Lemmatization** process aims to transform every word into the respective lemma. Lemma is the base form of a word that can be consulted in a dictionary (e.g., *"better"* → *"good"*). Lemmatization tools can reach accuracies of 97% in biomedical articles [130].

**Text chunking**, also known as shallow parsing, groups consecutive and syntactically correlated tokens into chunks, assigning labels to them. After the PoS tagging, chunking makes use of these tags to group words in higher order grammatical units, such as noun phrases (NPs), and verb phrases (VPs). A study of 6 chunkers for the biomedical domain using the GENIA corpus [121] obtained the F-scores, around 90% for NP chunking and 96% for VP chunking [131].

**Dependency parsing** is concerned with the analysis of the sentence structure, focusing on the relations between the words. A binary asymmetric relation between two tokens is called a dependency. An analysis of the outputs of several dependency parsers scored top accuracies of 90% [120].

### 2.3.3 Concept recognition

Concept recognition is the task that aims to automatically extract concepts (e.g. person names, diseases names, ...) from the text. Usually, the concept recognition task involves three main subtasks: Named Entity Recognition (NER), Normalization and Word Sense Disambiguation (WSD).

Named Entity Recognition (NER) aims to identify chunks of text and associate them with their specific concept type. This task can be performed by different approaches, such as dictionary matching [132], rule-based [133] or machine learning solutions [134]. Generally, the feasibility of each approach depends on the linguistic characteristics of the concepts being identified. Applying one of the described techniques, it is possible to automatically extract biomedical names from a massive amount of information. A brief comparison of some available NER tools is available in [8]. According to the authors, the best performing solution can achieve values of F-measure around 92% and 95%.

The goal of normalization and Word Sense Disambiguation (WSD) is to attribute each identified chunk of text to a unique concept from a curated knowledge base. This procedure is performed by associating a unique concept identifier, from databases or ontologies with each previously recognized concept. For instance, *"Alphaproteobacteria"* is an organism that is defined by the MeSH database, and its unique identifier is *"D020561"*. To apply this type of association, the creation of standards and ontologies for concept name definitions plays an important role in concept recognition tasks. The normalization task starts by associating the recognized name with any name on the biomedical knowledge scope. If there is no associated identifier, there is no option to assign an identifier to this concept, so it may be discarded as an entity name but if there is only one identifier associated, it is immediately endorsed. Otherwise, the entity name can be associated with more than one identifier. If this case occurs, it is considered ambiguous. Due to the biomedical domain complexity and extensibility, ambiguity is usual. For example, the term "cold" could refer to the temperature or a virus, depending on the context. Dealing with biomedical ambiguity problems is essential in order to achieve the correct identification of the concept names. To solve ambiguity, most of the WSD systems use machine learning techniques [135] and established knowledge solutions [136].

Figure 2.6 shows a recognition example, in which biomedical entities are identified and linked to curated resources, e.g. the MeSH database [44].

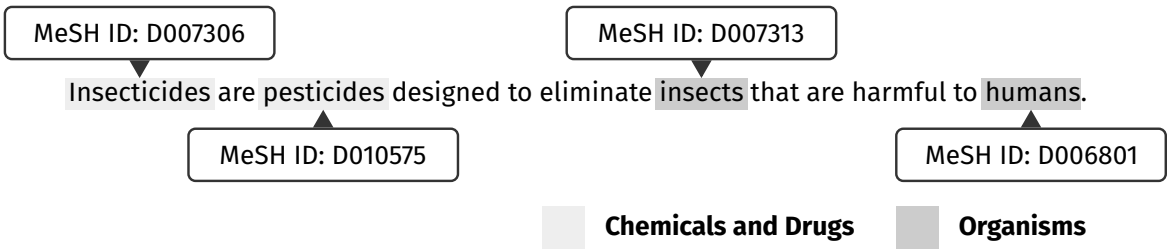


Figure 2.6: A recognition example of biomedical entities linked to the MeSH database [44] with their unique identifiers.

### 2.3.4 Relation mining

Biological systems involve interactions between entities, such as gene transcription or protein binding, elucidating the roles that biomolecules play in the biological processes. These relationships are usually described in the literature and are a good challenge for text-mining systems that apply relation mining techniques to extract and classify the

context and type of such relationships. The recognition of these interactions is an important step in biomedical information extraction [137], identifying diverse biological associations and involved entities.

Traditional relation mining solutions are focused on investigating and extracting direct associations between two concepts (e.g. genes, proteins, drugs, etc.) [138]. The study of these associations has generated much interest, especially in relation to protein-protein interactions (PPIs) [139], drug-drug interactions (DDIs) [140], and relations between chemicals and target genes [141]. As a consequence, the growing attention allied with the complexity of biological processes has generated new strategies to detect even more multifaceted interactions from text. Such complex interactions are typically distinguished in the literature as events.

Event extraction techniques became more common with the introduction of BioNLP shared tasks [142], a community-wide trend in text-mining for biology. In general, event mining aims to extract not only relations between concepts, but also relations between concepts and another relations, and even relations between relations allowing the construction of complex conceptual networks. Relation extraction is typically referred to as the task of extracting binary relations between concepts, and the event extraction as complex relation extraction involving verbs or normalized verbs to characterize the event type (i.e. trigger).

Figure 2.7 depicts a representation of a common relation extraction process between entities (a) and the respective event extraction process (b) of the same sentence. On the one hand, a) detects the protein and associates normalized relations (location) with the cells' components. On the other hand, in b) the localization event (translocation) captures the identification of the target (theme) entity (p65), the source (cytoplasm) and the destination (nucleus).

Overall, event representations capture the association of multiple participants with variable semantic roles [122] determined by domain requirements. Due to the complexity of the extraction process, most solutions only identify events and relations in a single sentence and not across sentences or papers.

In terms of performance, simple rule-based approaches can achieve an F-measure of 49% for biomedical event extraction [143]. For medical relation extraction, they can already achieve an F-measure of 67% [144].

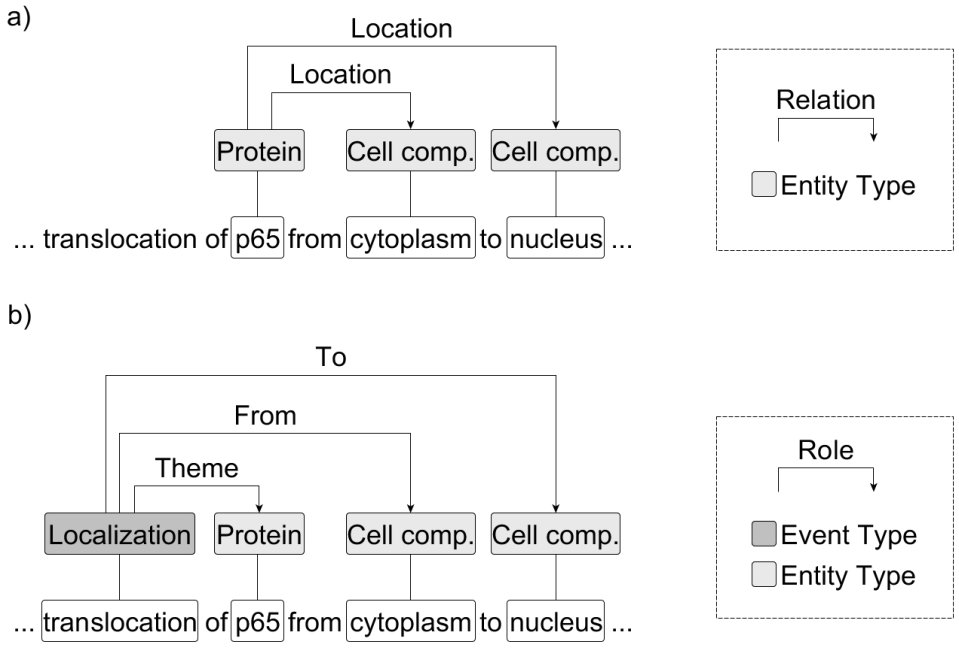


Figure 2.7: Analysis of the same sentence for relation and event extraction techniques.

### 2.3.5 Annotation formats

The results of text-mining solutions are typically kept in text files, using distinct data formats. Commonly called *annotations*, they are generated following a specific structure dependent on the extraction system.

#### Diversity

Several data formats have been proposed to represent biomedical information extraction outcomes. *IeXML* [145] was one of the first XML-based implementations to define an exchange format for annotations. More recently, *BioC* [146] has emerged as a community-supported format for encoding and sharing textual annotations. This simplified approach streamlines data reuse and sharing methods, achieving interoperability for the different text processing tasks. Figure 2.8 shows a *BioC* file extraction, containing a sample annotation of the *Alzheimer Disease* concept recognition. Although *BioC* provides interoperability between text-mined components, it is still a verbose format, not being designed to support data exploration and sustainability.

```
<?xml version="1.0" encoding="UTF-8">
<!DOCTYPE collection SYSTEM "BioC.dtd">

<collection>
...
 <document>
  <pmid>25766617</pmid>
  <passage>
   <text>Disturbances in the sleep-wake cycle
    and circadian rhythms are common
    symptoms of Alzheimer Disease...</text>
   <annotation id ="T1">
    <infon key="type">Disease</infon>
    <location offset="83" length="17" />
    <text>Alzheimer Disease</text>
    <id>OMIM:104300</id>
   </annotation>
   ...
  </passage>
  ...
 </document>
...
</collection>
```

File: 25766617.xml

Figure 2.8: BioC file extraction showing an annotation of the *Alzheimer Disease* concept.

Another possible organization for textual annotations is the *Standoff* format, in which annotations are stored separately from the annotated document text (Figure 2.9). For each text document, there is a corresponding annotation file. The two are associated with the file name. However, this simplified organization also hinders its wider use, namely if wanting to analyze the associations between the extracted data. Although there are several formats to represent biomedical textual annotations, its organization in a SW compliant format allows easier exploration of this information.

### Integration

Emerging SW standards and concepts are currently seen as the standard paradigm for data integration and distribution on a web-scale, focused on the semantics and the context of data [147]. It allows the construction of rich networks of linked data, offering advanced possibilities to retrieve and discover knowledge (e.g. reasoning). With the

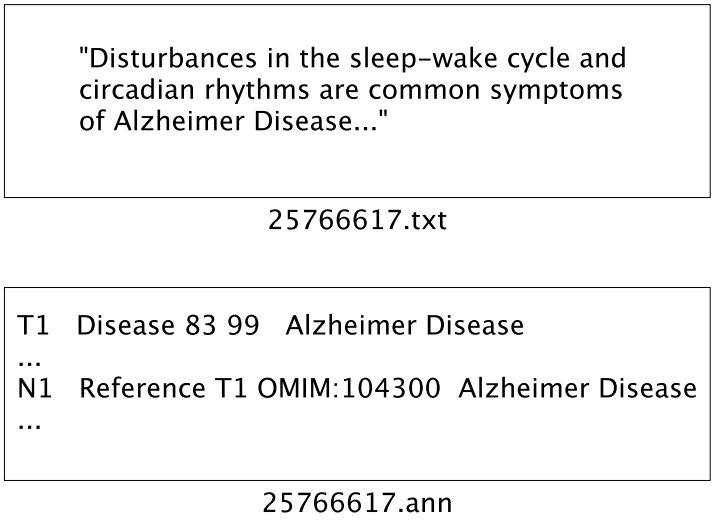


Figure 2.9: Standoff file extraction showing an annotation of the *Alzheimer Disease* concept.

increasing adoption of this paradigm to tackle traditional data issues such as heterogeneity, distribution and interoperability, novel knowledge-based databases and systems have been built to explore this technology's potential. Essentially, they facilitate the deployment of well-structured data and deliver information in a usable structure for further analysis and reuse.

In this way, approaches combining the benefits of information extraction methods with these semantic systems represent a growing trend, allowing the establishment of curated databases with improved availability [148]. Coulet *et al.* [149] provide an overview of such solutions, and describe a use case regarding the integration of heterogeneous text-mined pharmacogenomics relationships on the SW. Another case study is described by Mendes *et al.* [150], presenting a translation method for automated annotation of text documents to the DBpedia Knowledge Base [82]. The developed strategy represents ongoing efforts striving to integrate the existing knowledge in text documents into the Linked Open Data network. A different approach is proposed through the PubAnnotation [151] prototype repository. The notion was to construct a sharable store, where several corpora and annotations can be stored together and queried through SPARQL.

In this perspective, there is a clear trend to combine text-mined information with SW technologies, resulting in improved knowledge exchange and representation. Taking into account these approaches, there is a clear tendency towards workflow construction systems for annotation distribution. However, limitations in the development processes and the

existence of software dependencies in the source platforms [152] represent a barrier to adapting and reusing existing solutions for the distribution of distinct annotation structures and formats. The great heterogeneity of biomedical annotations makes it challenging to aggregate results obtained from different tools and systems, with innovative solutions being necessary for the combination and distribution of multiple annotations.

## 2.4 Representation of scientific knowledge

The continuous increase of biomedical data has provided good opportunities for scientific achievements. Yet, at the same time, its management is challenging for the research community. Information from studies needs to be adequately processed and shared between stakeholders [153] to enable an incremental understanding of biological and medical processes. Although several methods are used, the most common is the publication of scientific articles in international peer-reviewed journals. Traditionally, the evaluation of researchers' scientific output is based on this process. Over the years, this methodology has persisted mostly due to its ability to attribute credit to the authors. However, it is still unclear how academic credit is established for biomedical data sharing, and traditional journals can provide semi-structured information that is coherent and a valuable addition to scientific knowledge.

Furthermore, we cannot exclude existing methods that also contribute to the scientific field. Examples of these include the submission to, or curation of, biological databases [154]. In these particular cases, there is no effective way to credit authors' work.

Additionally, the biomedical field has faced concerns due to an excessive amount of information [155]. Most of this material is cumulative and finding the desired data and related connections, including provenance details, requires considerable efforts. To tackle these challenges, novel methodologies are needed to effectively summarize and credit scientific knowledge.

### 2.4.1 Available strategies

One of the earliest practical demonstrations of the "microattribution" concept occurred in 2011 with the publishing of a set of locus-specific databases for publishing genetic variation related data [156]. From that moment, the concept itself as an alternative reward method for scientific contributions has generated a lot of interest in the scientific community.

With the dawn of the SW paradigm [11], the nanopublication notion [157] has emerged to solve most data credit issues. Several prototypes have been developed using the nanopublication format to obtain credit for the shared content. These cover several areas, including public and commercial drug discovery research [158], human gene-disease associations [159], human proteins [160] and specific rare diseases [161].

Sample first-generation nanopublications were produced from the Leiden Open-Access Variation Database (http://www.lovd.nl) [162], encouraging the submission of human genomic variant data for sharing within the scientific community [154]. A more generic conversion approach is provided by Prizms [163]. According to the authors, the tool converts several formats into a nanopublication model called "datapub", to describe the integrated datasets. A proof-of-concept demo with 330 melanoma datasets is available at `http://data.melagrid.org`. A different and interactive approach is provided by the Nanobrowser portal (`http://nanobrowser.inn.ac`). Nanopublications are created through a manual process using sentences to characterize underspecified scientific claims [164]. The sentences are built using the AIDA (Atomic, Independent, Declarative, Absolute) semantic scheme providing a similar and summarized representation of scientific assertions [165].

### 2.4.2 Nanopublications

Nanopublications allow authors to interconnect and exchange data while receiving credit for shared content. The idea is that they are more suited than traditional papers to represent the relationships that exist between research data, providing an efficient mechanism for knowledge exchange [166]. They are built through SW concepts and strategies, allowing knowledge summarization with proper attribution mechanisms. This normalizes how provenance, authorship, and related publication information can be attributed, reusing information whenever possible. Serializable on the RDF interoperable format, it facilitates knowledge exchange methodologies fostering retrieval and use. Furthermore, it provides the possibility of being cited and the impact tracked with the universal nanopublication identifiers, boosting compliance with open SW standards.

#### Model

Currently, the nanopublication community (`http://nanopub.org`) is developing this standard through an incremental process. Several efforts are underway to produce

guidelines and recommendations for the ultimate schema (`http://nanopub.org/nschema`).

Figure 2.10 presents the fundamental structure according to the nanopublication model. The basic nanopublication structure comprises four main components, in which the unique nanopublication *Identifier* is associated with the *Assertion*, *Provenance* and *Publication Information*. Both fields comprise a set of RDF triples describing the nanopublication metadata. Regarding the *Assertion* graph, one assertion at least must be included to be valid: an assertion is the smallest unit of thought, expressing a relationship between two concepts. Supporting metadata about the assertion context are provided through the *Provenance* graph. This includes methods that were used to generate the assertion and attribution metadata such as, DOIs, URLs, timestamps, authors and related information. Supplementary metadata about the nanopublication itself is enclosed in the *Publication Information* graph. Attribution, generated time, keywords or tags are sample metadata that can be added to offer provenance information regarding the nanopublication itself.

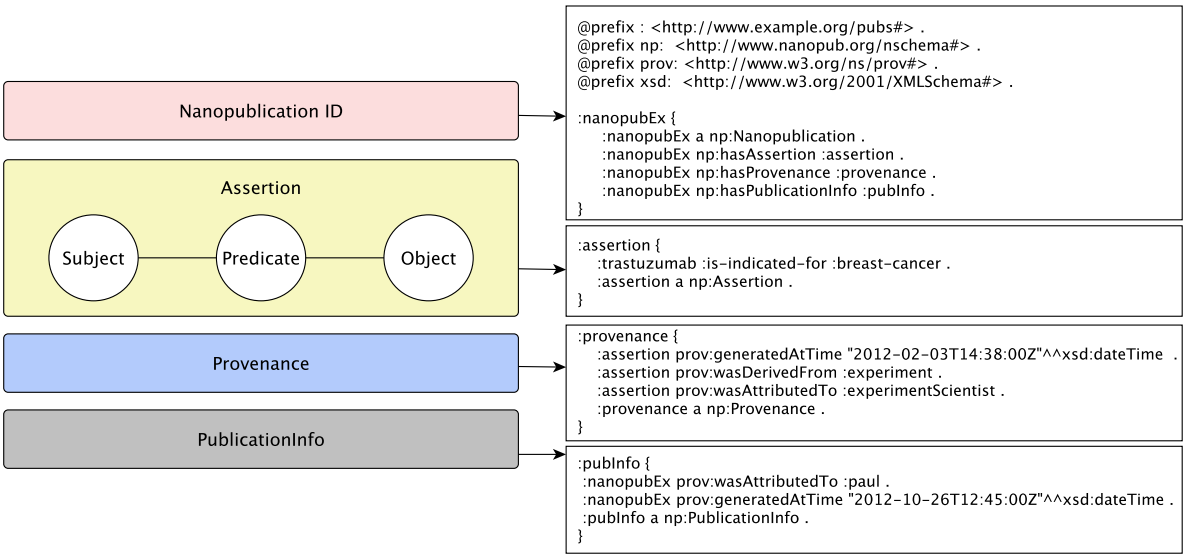


Figure 2.10: Anatomy of a nanopublication. The different fields of the nanopublication structure (left) with corresponding example (right) in TriG syntax (`http://www.w3.org/TR/trig/`).

### Opportunities

The nanopublication strategy offers not only a great opportunity to improve and publish conventional paper research data, but also to explore positive and negative data studies.

Experimental data studies can also be published in a standard format, such as RDF triples, instead of archived as supplementary information in an arbitrary format or independent databases. Research Objects (RO) [167] are a possibility for the aggregation objects that bundle experimental studies' resources. However, in contrast to nanopublications, RO encapsulate elements for an entire investigation, as opposed to individual claims [168]. With the nanopublication strategy, researchers can access more quickly not only the more relevant information but also the supporting data and related metadata. It also intends to reduce the publishing time and the search period, streamlining the investigation procedure. For these reasons, deploying data as nanopublications will improve similar studies, saving time and unnecessary costs. In a sense, they are a natural response to the increasing number and complexity behind scientific communications. As a result, the aim is to overcome the inconsistency, ambiguity and redundancy of classical publications, enhancing information extraction and analysis.

However, even with the adoption of standards, some data sharing issues persist. The major reason is the lack of expertise to convert local data into accepted data standards [163]. The inexistence of adequate tools and systems are obstacles to data-publishing as nanopublications. Current solutions are developed for specific scenarios and based on hand-scripted processes, which limits their applicability.

In this way, they fail to address three important features. First, they are produced by hand-scripted processes, the main functionality being missing: an automated transformation from several legacy data formats to nanopublications, following the latest schema (`http://nanopub.org/nschema`). Second, they do not follow an interactive approach. Cooperative solutions that are easy to set up by non-informaticians and able to effortlessly perform this transition are crucial to overcome the limited deployment of nanopublications. Third, they are based on specific use cases, with reproduction or adaptation being unfeasible. Therefore, researchers clearly need an easy set-up procedure that allows them to publish and share their scientific outcomes through a reliable system [26].

## 2.5 Summary

SW is becoming a common bridge across biomedical silos of disconnected standards. The true value behind this technology lies in how easy it is to access and exchange knowledge between independent systems toward making software easier to connect in the

future. In this way, a shift of traditional resources to this modern paradigm is vital to better promote, express and infer biomedical knowledge.

Despite several efforts to drive biomedical resources into the SW ecosystem, there is still a need to provide seamless integration of diverse information such as omic data, scientific literature and many other data sources (e.g. databases, spreadsheets, etc.).

Following the problems highlighted in the current state-of-the-art, four solutions will be presented and discussed to enable interoperability across biomedical resources. First, the connection of distributed resources is addressed. Second, the integration of semi-structured data formats is investigated. Third, the semantic integration will be mainly focused on unstructured data. Finally, a higher level of semantic integration will be discussed for fast deployment of new, modern and high-performance information systems.

# Chapter 3

# Connecting rare disease patient registries

Patient registries are an essential tool to increase current knowledge of rare diseases. Understanding these data is a vital step to improve patient treatments, and to create the most adequate tools for personalized medicine. However, the growing number of disease-specific patient registries also brings new technical challenges. Usually, these systems are developed as closed data silos, with independent formats and models, lacking comprehensive mechanisms to enable data sharing. To tackle these challenges, we developed a semantic web-based solution that allows connecting distributed and heterogeneous registries, enabling the federation of knowledge between multiple independent environments[1]. This semantic layer creates a holistic view over a set of anonymized registries, supporting semantic data representation, integrated access and querying. The implemented system gave us the opportunity to answer challenging questions across disperse rare disease patient registries. The interconnection between those registries using Semantic Web technologies benefits our final solution in the way that we can query single or multiple instances according to our needs. A web-based entry point is available at `http://bioinformatics.ua.pt/linked-registries-app/`. This strategy allows a holistic view through connected registries, enabling state-of-the-art semantic data sharing and access. The outcome is a unique semantic layer, connecting miscellaneous registries and delivering a lightweight holistic perspective over the wealth of knowledge stemming from linked rare disease patient registries.

[1] This chapter is largely based on the paper by P. Sernadela, L. González-Castro, C. Carta, E. van der Horst, P. Lopes, *et al.*, "Linked Registries: Connecting Rare Diseases Patient Registries through a Semantic Web Layer", *BioMed Research International*, vol. 2017, pp. 1–13, Oct. 2017.

## 3.1 Overview

Rare disease patient registries are typically fragmented by data type and disease. Furthermore, these systems have poor interoperability due to the high complexity and heterogeneity of data types and the lack of standards in data models and data descriptions. Due to strict data-protection requirements, access to patient registries is restricted, converting these valuable, distributed sources in closed data silos. This is a barrier to linking patient-centric electronic records across registries and diseases. To tackle these barriers to integration, we started the study by looking for relevant questions difficult to answer without an infrastructure of integrated patient data from several registries. The motivating questions were, for instance:

1. Given a set of phenotypes that are relevant for neuromuscular (e.g. DM, FSHD, LGMD2I) and neurodegenerative diseases (e.g. HD, Ataxia), can we find patients in a disease non-specific way?

2. More specifically, based on "Ambulation", "Age", and "Country", can we get the number of patients?

Answering this kind of question requires data of a disparate nature and from multiple sources. To harmonize these data into a semantic layer we need an integration platform capable of converting any data format into RDF. This implies abstracting registry concepts and their attributes, such as: Patient (sex, date of birth, country); Disease; Phenotype (motor, ambulation); and Genetic Variation, and then, representing them in a graph data model in which the semantics of the objects and their relationships can be described with standard or widely adopted ontologies. Finally, patient registry data should be aggregated by concept. In each concept, data elements, or instances that represent the same entity but have different text mentions in each registry, must be mapped to an ontology term. For example, Orphanet Rare Disease Ontology (ORDO) for diseases, and the Human Phenotype Ontology (HPO) for phenotypes can be used to make data interoperable and linkable in the Web of data. The use of domain-specific and commonly used ontologies adds value to data, through an integrated knowledge base that is searchable and comparable by users and machines [169]. Furthermore, interlinking patient registry data with external linked datasets allows enrichment of current knowledge in rare disease research. In this work, we have developed a new semantic layer on top of existing patient registries, to allow extracting anonymised data from the original datasets, translating them

to a common shared exchange model and making them available to the research community (available at `http://bioinformatics.ua.pt/linked-registries-app/`). The solution addresses three key requirements from the patient registries research community: 1) data model agnostic; 2) distributed and encapsulated; and 3) knowledge-oriented. Firstly, data harmonization strategies are data model agnostic and work regardless of registries' data format and internal structure. This is clearly important as we are dealing with systems featuring assorted characteristics, from relational databases and service endpoints, up to Excel spreadsheets. Next, the solution is distributed and encapsulated. When dealing with rare disease patients, it is imperative to ensure data anonymity and privacy. Hence, we need tools that extract meaningful data while maintaining hidden all the attributes that may disclose patients' identification. Finally, our approach takes advantage of semantic web technologies to improve how we publish, access, express and share knowledge across the Web. From a technological perspective, the system was built on top of COEUS [20], an application framework that streamlines data integration with semantic representation. As patient registries are shared within this platform, researchers and developers are able to perform federated queries, covering miscellaneous databases, just as they would query a single local dataset. In summary, we explore a semantic web approach and a non-intrusive strategy to interconnect, enrich and federate data from multiple patient registries, allowing extension of the knowledge contained in these distributed repositories.

## 3.2 Architecture

Semantic data integration is, in itself, a complex data engineering issue, if we have to code every component of the software solution [170, 171]. Building on previous results [17], we use COEUS as the baseline framework of our platform. Exploring its flexible integration engine enables simplifying the overall platform architecture through the creation of a comprehensive dependency-based resource integration network. Figure 3.1 presents the platform's distributed architecture, which is organised in four levels: 1) Patient, 2) Semantic, 3) Federation, and 4) Research. At the patient level we gather information from the distributed and heterogeneous patient registries, which can be stored in multiple formats and using various technologies (e.g., relational databases, text files, spreadsheets, etc.). Patient registries can be integrated in the framework regardless of their location, and their quantity. At the second level we include additional semantics in patient registry data. This is done using COEUS, which acts as the main abstraction,

storage and publishing engine. Here, we manage the anonymised patient data, translating from their primitive format to common biomedical ontologies. The third level provides the knowledge federation and data exploration capabilities, i.e. SPARQL queries can be forwarded to several patient registry endpoints. COEUS acts here as a middleware component between the patient registry triplestore and the public knowledge federation layer. Finally, at the upper level researchers can perform general queries that combine data from one or more patient registries. In a sense, query federation enables performing SQL-like UNIONs or JOINs across multiple knowledge bases. This allows knowledge inference and reasoning queries to go beyond what is currently possible.

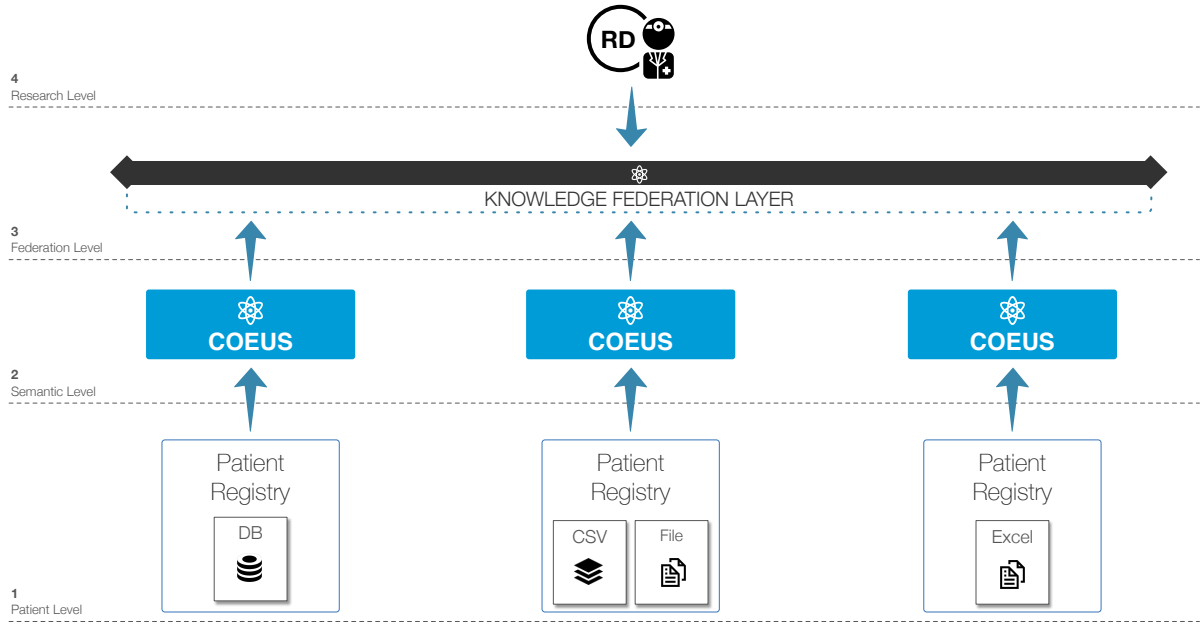


Figure 3.1: Knowledge federation architecture, integrating distributed patient registries via COEUS.

## 3.3 Workflow

Publishing anonymized patient registry data in a semantic way requires a comprehensive workflow. Figure 3.2 describes the key steps in this semantic integration and translation pipeline: 1) ontology mapping; 2) COEUS setup; 3) semantic translation and 4) data publishing. The first step consists of defining the best ontologies to map common patient

data. HPO [74], UMLS [172], ICD [173] or ORDO [65] are the most widely used ontologies in the field of rare diseases. One of the great advantages of using semantic web technologies is that any external ontology can be used to complement or extend COEUS internal model. As long as clinicians understand the new predicates, any number of properties can be included, semantically mapping concepts or entities to existing ontologies, or adding further properties to describe entities or concepts. Moreover, we may combine multiple ontologies, i.e., the same data element can be mapped to terms from more than one ontology, optimising its expressiveness and enriching the way it can be used in future research environments. In this step, semi-automated annotation tools such as SORTA [138] and EGAS [119] among human curation experts play an important role in the annotation of biomedical data (e.g. phenotypes, diseases, etc). The second step of the pipeline consists of the configuration and deployment of a new COEUS instance. The setup involves defining how data will be extracted and mapped in the selected ontology terms. Using COEUS connectors we have to specify where the data comes from (Excel, CSV or XML files; SQL databases; or SPARQL / Linked Data endpoints), and how we will map it to the ontologies. For instance, for a patient registry available as a CSV file, we need to specify the file location and, for each mapped ontology term, the column containing the actual data elements. In the following stage, the semantic translation process, knowledge base elements and their data and object properties are created in real-time from the integrated data. This step elevates data in primitive formats to a new semantic abstraction level. The process is complete when all data are imported into a new triplestore, making it available for external use through the various data publishing endpoints.

## 3.4 Implementation

COEUS framework is focused on helping researchers in the construction and publishing process of new semantically enhanced systems. It offers a good starting point to integrate disparate data due to the advanced ETL (Extract-Transform-Load) processes in its engine. These algorithms facilitate the "triplification" process, in which all data are converted to a simple subject-predicate-object model. Moreover, it makes the integrated information available through a hierarchical model establishing relationships between data in an *"Entity-Concept-Item"* structure (e.g. *Protein-Uniprot-P51587*). To create each registry's knowledge base according to this organized model, we must fulfill some initial requirements. Essentially, there are three main steps to achieve the final solution: the first is data

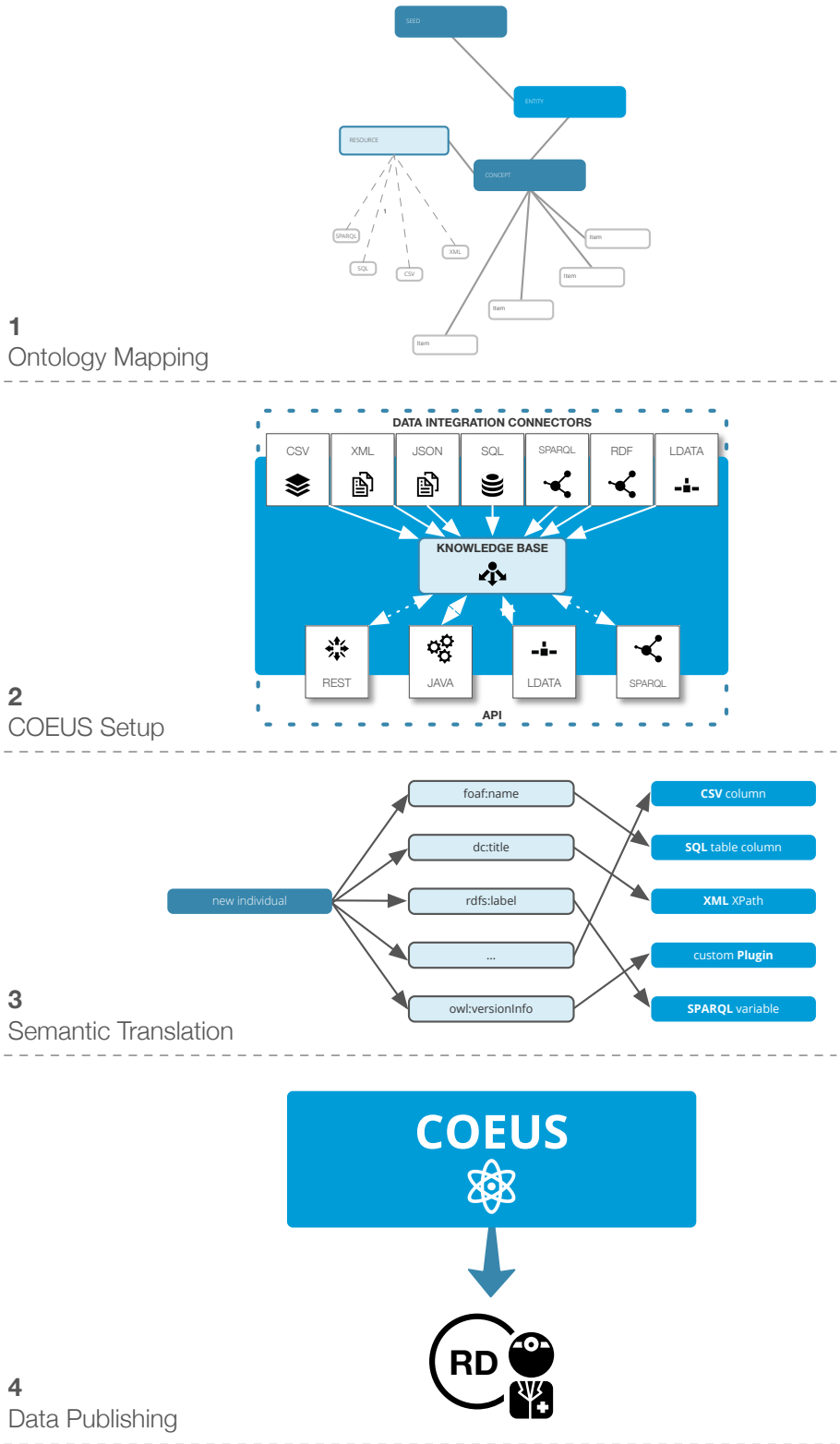


Figure 3.2: Simplified registry publication workflow.

selection, the second is the data sources' configuration and last one is the data integration process.

In the first step, we have applied a pre-selection process to all registries' data. In this process, we select the desired information for our study avoiding sensitive data and considering only non-identifiable data. Patient registries have a lot of information, but some is redundant and incomplete. Therefore, applying this initial filtering process is vital to build a consistent database. This process was only possible due to the involvement of respective data owners who translated what their data mean. In the final stage of this initial filtering process, we exported the files of the distinct registries. Each dataset was imported to a separated COEUS instance, as we plan to have four distributed systems.

In the second step, we define the data sources' attributes for each registry. This starts by creating one *Resource* (Figure 3.2, block 1 – Ontology Mapping) in the knowledge base that contains the respective data elements such as Endpoint location (i.e., the registries' file location), Publisher (i.e., CSV file), CSV Starting line (i.e., *1* as the registry file has headers) and Method (i.e. *cache* as we will load the entire file onto the system). Additionally, for each *Resource*, a combination of parameters (i.e., *Selectors*) must also be included to establish the mapping between the information to be extracted from the registry (e.g. for each column) and the respective formal ontology terms that connect it. For instance, we can make use of the Human Disease Ontology [90] term *doid:has_symptom* to establish the connection between a Facioscapulohumeral muscular dystrophy (FSHD) patient and the identified symptoms: *coeus:Patient_X doid:has_symptom obo:HP_0001324* (i.e. Patient X has a muscle weakness symptom). Likewise, we link each patient to its respective identifier by creating a *Selector* that makes the linkage between the CSV first column (with the patient IDs) in the parameter query and the property *dc:identifier* from the Dublin Core Ontology [174]. Establishing all these mappings in the registries' records allows the foundation of an interconnected network of relationships between patients and respective features. An overview of the knowledge base model is available in Figure 3.3, showing a simplified view of Facioscapulohumeral Muscular Dystrophy Type 1 (*omim:158900*) patients' relationships.

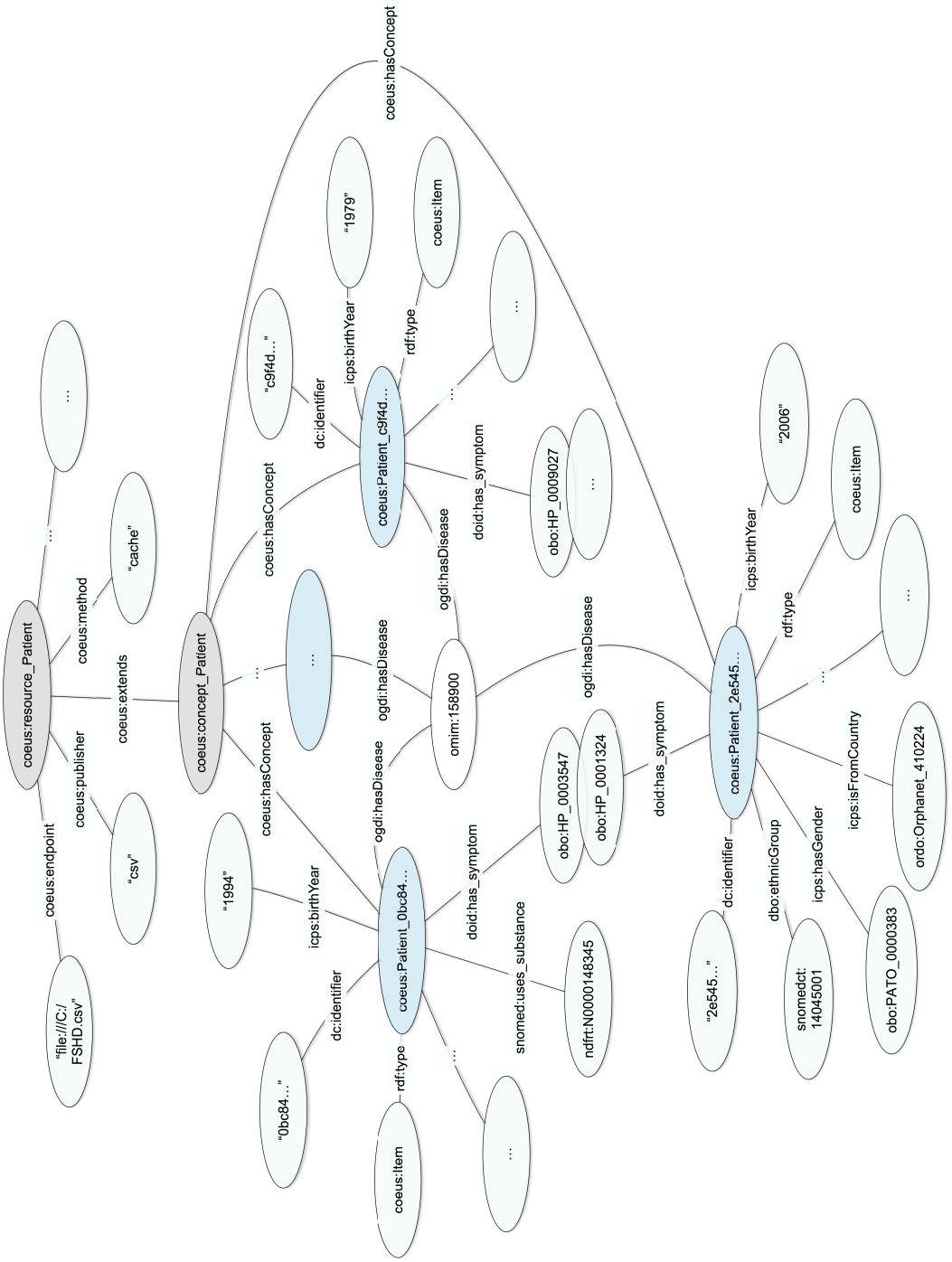


Figure 3.3: Patient registry model overview. Facioscapulohumeral Muscular Dystrophy patients share concepts and relationships, creating a fully-connected network.

After these configurations, the last process encompasses the automatic integration and semantic mapping of data sources. To expose data, we provide several interoperability features such as SPARQL endpoint and a Linked Data interface. The SPARQL endpoint works as a federated query system, in which we can perform complex queries across patient registries. The Linked Data interface provides easy access to patient information through the Web Browser or similar applications. All the processes described are managed through COEUS Web user interface, which provides an easy set-up solution for the installation and configuration process.

## 3.5 Results

Neurodegenerative and neuromuscular diseases are amongst the most frequent rare diseases, affecting the life and mobility of more than 500,000 patients and families in Europe (`http://rd-neuromics.eu/`). We used the proposed architecture to integrate four patient registries in the neuromuscular and neurodegenerative disease area. These registries collected patient data from ten different countries (United Kingdom, Italy, Spain, Denmark, France, Netherlands, Sweden, Austria, United States and Germany), and information related to four rare diseases: Myotonic Dystrophies (DM), Facioscapularhumeral muscular dystrophy (FSHD), Fukutin Related Protein (FKRP) related conditions (e.g., LGMD2I) and Huntington's Disease (HD).

### 3.5.1 Exploring rare disease patient registries

To guide the development of our solution and also, at the end, to allow its validation, several questions were initially elaborated:

1. Can we find more than ten males and ten female patients that share a set of phenotypes and live in different countries?

2. To build a trial/research data set, how many patients have the desired conditions/requirements for the study and live closest to the clinical/research setting?

3. What are the phenotypes associated with Myotonic Dystrophies (DM) and Facioscapularhumeral muscular dystrophy (FSHD) diseases?

4. Are there patients treated with different therapies diagnosed with the same disease?

5. Are there patients treated with the same therapy but diagnosed with different diseases?

6. Can we find patients diagnosed with a certain disease with different states of morbidity?

7. Are there patients with this specific set of phenotypes?

8. Are there patients sharing phenotypes and diagnosed with different neuromuscular and neurodegenerative diseases?

To find answers to these questions researchers and clinicians need to look for patient data that are fragmented in different registries and combine data across registries and across diseases. Without an infrastructure of integrated patient registries, this is not straightforward because each patient registry is designed, described and technically implemented in a particular way, and data are not connected. This means that to answer any of these questions a considerable amount of time will be spent understanding each registry data model, to access and retrieve the necessary data from each registry, aggregate all registries' data in a meaningful way, and finally, to query-answer over the harmonized data to extract the information. These are rather inefficient, impractical propositions. Thus, to gain a complete view of a specific disease and patient population of interest, and to retrieve the desired information to answer these questions, the linking of registries' data sets is an essential step. The described research questions involved the collaboration of data owners and database managers, who participated actively in the data selection and harmonization processes, and provided continuous feedback for the final solution.

### 3.5.2 The Linked Registries solution

The interconnection between disperse patient registries using COEUS facilitates our final solution in the way that we can query single or multiple instances according to our needs. However, to better access and enhance user interaction, we provide a single, web-based entry point to access the aggregated information available in each instance.

The entry point is available at `http://bioinformatics.ua.pt/linked-registries-app/` and is based on a combination of SPARQL federation queries templates with pre-defined variables. In each template question, variables can be adjusted according to the knowledge base values automatically. All queries are generated in real-time by the application and can be edited or adjusted (by using the advanced mode) for more

accurate examination. Using this solution, users have the opportunity to find answers to the previously defined questions across disperse patient registries. For instance, to answer the first question we created a template federated query to retrieve results from all the registries. As we store countries' information in each COEUS instance using the same model, we are able to retrieve common characteristics from each instance. Therefore, finding a cohort sharing a set of phenotype information and based in different countries can be a straightforward process (Figure 3.4).

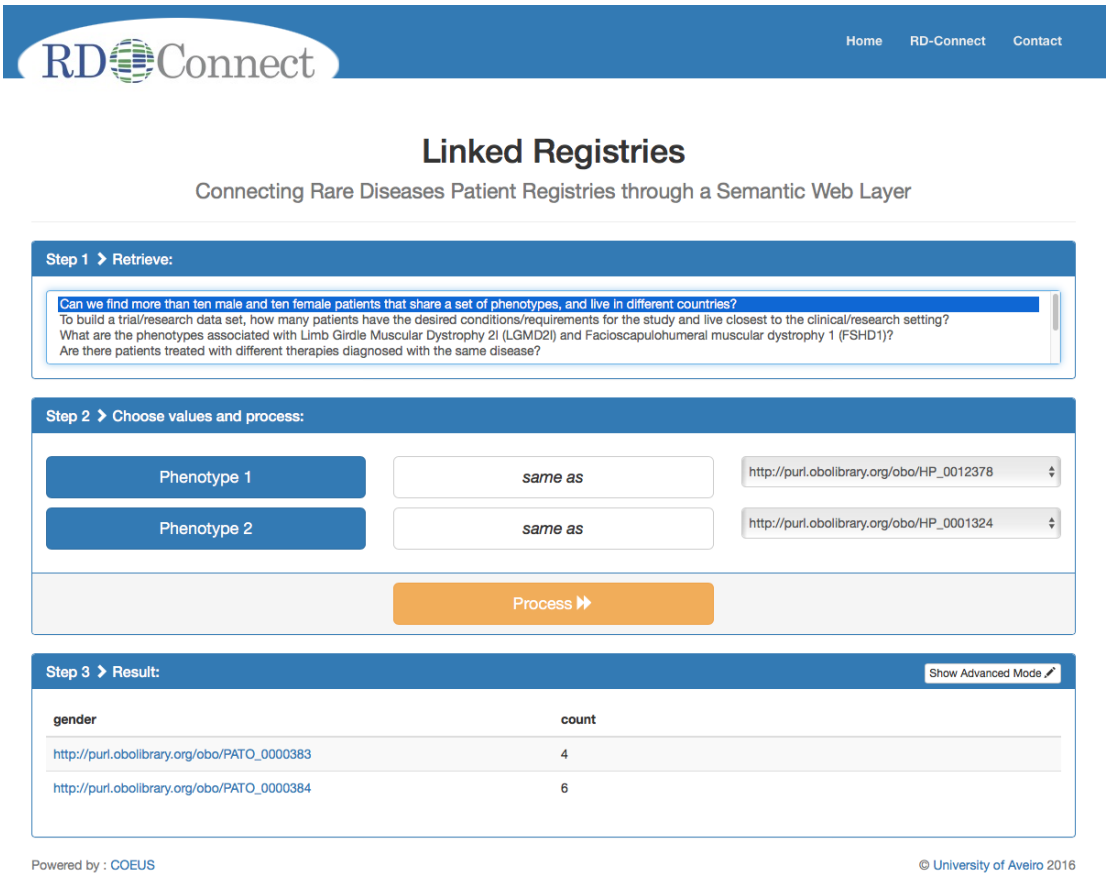


Figure 3.4: Linked Registries web application interface.

In order to answer this type of question, we searched for male and female patients that have both "fatigue" (i.e. *obo:HP_0012378*) and "muscle weakness" (i.e. *obo:HP_0001324*) phenotypes. Querying our system, we retrieved male patients from six different countries, and female patients from four. Therefore, for our particular set of registers, the number of patients (either male or female) living in different countries and sharing those phenotypes

is under ten. Concerning question 2, filtering patient characteristics according to some conditions, such as the use of a gastric tube on a Myotonic dystrophy type 1 (MD1) patient, can be successfully performed. However, to discover patients living closest to their clinical/research setting is not a trivial task to perform due to the limitation of our registries' data, which only covers country-related information. It is also possible to answer enquiries similar to question 3. For instance, we can query the two remote databases (e.g. FKRP and FSHD) through their SPARQL endpoint and search for all shared phenotypes that have been registered for patients suffering from Limb Girdle Muscular Dystrophy 2I (LGMD2I) and Facioscapulohumeral muscular dystrophy (FSHD) diseases to answer this question. The SPARQL query is as follows:

```
PREFIX doid: <http://purl.obolibrary.org/obo/doid#>
PREFIX ogdi: <http://purl.bioontology.org/ontology/OGDI#>
PREFIX omim: <http://purl.bioontology.org/ontology/OMIM/>

SELECT DISTINCT ?phenotype {

    SERVICE <FKRP-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient_FKRP ogdi:hasDisease omim:607155 .
        ?patient_FKRP doid:has_symptom ?phenotype
    }

    SERVICE <FSHD-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient_FSHD ogdi:hasDisease omim:158900 .
        ?patient_FSHD doid:has_symptom ?phenotype
    }

    FILTER (isURI(?phenotype))
}
```

In our case, the result of this query returned the respective shared phenotypes for both diseases: "fatigue" (*obo:HP_0012378*), "muscle weakness" (*obo:HP_0001324*) and "rigidity" (*obo:HP_0002063*). The same occurs for questions 4) and 5) as we have collected information regarding patients' diagnosis. This information was integrated according to the specifications for each disease. However, we are able to cross information between therapies and diseases due to our standardization strategy based on community-shared

and common ontologies in all patient registries. For instance, if we query the different registries looking for patients treated with "ACE inhibitors" (i.e. *ndfrt:N0000029130*), we can easily find a correlation between DM (*omim:160900*) and LGMD2I (*omim:607155*) diseases:

```
PREFIX snomedct: <http://purl.bioontology.org/ontology/SNOMEDCT/>
PREFIX ogdi: <http://purl.bioontology.org/ontology/OGDI#>
PREFIX ndfrt: <http://purl.bioontology.org/ontology/NDFRT/>

SELECT DISTINCT ?disease WHERE {
{
    SERVICE <EHDN-REGISTRY-SPARQL-ENDPOINT>
    {
     ?patient_EHDN snomedct:uses_substance ndfrt:N0000029130 .
     ?patient_EHDN ogdi:hasDisease ?disease
    }
} UNION {
    SERVICE <DM-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient_DM snomedct:uses_substance ndfrt:N0000029130 .
        ?patient_DM ogdi:hasDisease ?disease
    }
} UNION {
    SERVICE <FKRP-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient_FKRP snomedct:uses_substance ndfrt:N0000029130 .
        ?patient_FKRP ogdi:hasDisease ?disease
    }
} UNION {
    SERVICE <FSHD-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient_FSHD snomedct:uses_substance ndfrt:N0000029130 .
        ?patient_FSHD ogdi:hasDisease ?disease
    }
}}
```

However, answering questions such as 6 is very complex. The difficulty resides in finding structured patient states of morbidity in each registry. Disease states are usually stored and described as long plain-text fields without suitable structure, which makes the task of finding similarities in that information more complex. Additionally, not all patient

registries have this type of information, creating barriers to the crossing of information among different registries. Therefore, we do not integrate different states of morbidity of each patient into our system. In contrast, the two following questions, 7 and 8, can be more easily answered due to the structured information available about different diseases and respective patient phenotypes. To give an example for question 7, we can randomly choose phenotypes such as "fatigue" (*obo:HP_0012378*) and "muscle weakness" (*obo:HP_0001324*), and simply count how many patients share both:

```
PREFIX doid: <http://purl.obolibrary.org/obo/doid#>
PREFIX obo: <http://purl.obolibrary.org/obo/>

SELECT (COUNT(DISTINCT ?patient) as ?count) WHERE {
{
    SERVICE <EHDN-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom obo:HP_0012378 .
        ?patient doid:has_symptom obo:HP_0001324
    }
    } UNION {SERVICE <DM-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom obo:HP_0012378 .
        ?patient doid:has_symptom obo:HP_0001324
    }
    } UNION {SERVICE <FKRP-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom obo:HP_0012378 .
        ?patient doid:has_symptom obo:HP_0001324
    }
    } UNION {SERVICE <FSHD-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom obo:HP_0012378 .
        ?patient doid:has_symptom obo:HP_0001324
    }
}}
```

Using this schema allows us to find up to forty-one patients spread over the different databases. To answer question 8, we can also make a federated query to all registries to retrieve a list of associations between phenotypes and diseases. Therefore, we are able to detect the most common associations by counting the number of patient occurrences in

which the phenotype-disease association was identified:

```
PREFIX doid: <http://purl.obolibrary.org/obo/doid#>
PREFIX ogdi: <http://purl.bioontology.org/ontology/OGDI#>

SELECT ?phen ?disease (COUNT(DISTINCT ?patient) as ?count) WHERE {
{
    SERVICE <EHDN-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom ?phen .
        ?patient ogdi:hasDisease ?disease
    }
    } UNION {SERVICE <DM-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom ?phen .
        ?patient ogdi:hasDisease ?disease
    }
    } UNION {SERVICE <FKRP-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom ?phen .
        ?patient ogdi:hasDisease ?disease
    }
    } UNION {SERVICE <FSHD-REGISTRY-SPARQL-ENDPOINT>
    {
        ?patient doid:has_symptom ?phen .
        ?patient ogdi:hasDisease ?disease
    }
}
    FILTER (isURI(?phen) )
    FILTER (isURI(?disease) )
}
GROUP BY ?phen ?disease ORDER BY ?count
```

By querying our system, we are able to detect, for instance, that phenotypes such as "muscular weakness" (*hp:0001324*) and "fatigue" (*hp:0012378*) are more common in muscular dystrophy (*omim:607155*) diseases and phenotypes such as "Myotonia" (*hp:0002486*) and "fatigue" (*hp:0012378*) are more representative in myotonic dystrophy type 1 (*omim:160900*) diseases.

## 3.6 Discussion

The International Rare Diseases Research (IRDiRC) consortium defined several overarching objectives, to achieve by 2020 [175]. Some of these goals include, for instance, making data accessible to the research community, or promoting tools and standards that simplify networking between data centres. The present solution was built upon these general needs, offering an opportunity to access patients' distributed data in a common web platform. The semantic layer approach offers a technological solution that enables data and metadata sharing, following common ontologies and standards, as described throughout the document. In our research, we identified how these semantic web technologies can be tailored to the patient registry integration scenario. Although our results are successful, they highlight two major issues.

First, identifying the proper common ontology to be used across patient registries is a cumbersome task. While COEUS allows this process at the technical level, there still has to be an agreement between stakeholders on what ontologies will be used and how their data will be properly mapped to them. This introduces a new challenge, as distinct ontologies need to be adequately mapped [176]. At this point, it is important to highlight that the creation of mappings between patient registry elements and ontologies is a critical point for data quality and reliability. Rare disease registry researchers frequently need to extract the primary clinical information and translate it into the registry data elements. This process is the key to the validity of outcomes in the scope of the registry. Both phenotypic information and final diagnosis have to be derived from the clinical examination, genetic, histo-pathological and other laboratory tests and radiological images, among some other specific sources, which are all challenging due to their heterogeneity and complexity. In addition, standardization of the primary sources of information is an important issue for registries, but in some situations, it is not possible. The translational process from the real clinical status of the patient to the information saved and stored in the registry database implies a potential risk of introducing some biased information. The establishment of mappings between information based on ontological terms could lead us to obtain standardized data, but not valid results. Thus, when phenotypic data are not well defined or are incorrectly translated into the database elements, this phenotypic information might be linked to wrong ontological terms. Likewise, ontological terms are not always as comprehensive as free text, and therefore, the ability of an ontology to cover all phenotypic traits of specific diseases is another limiting factor. In this regard, collaboration is needed with ontology developers in order to expand with further

ontological terms and thereby align ontology representations with the current knowledge. This active translational dialogue among the actors in clinical and research domains is important to both stimulate the use of standards in patient registries and to ensure an appropriate description of the current domain knowledge in biomedical ontologies. In this challenging scenario, the mapping of clinical terms has to be undertaken according to quality procedures. Nevertheless, several organizations are publishing common data element models in order to solve the interoperability problem among different patient registries. Although these efforts ensure interoperability within the selected domain, interoperability across application domain boundaries is not automatically possible [177]. There are over 600 rare disease registries in Europe alone, the majority not currently using a specific ontology. Despite the overall desire in the community to increase harmonization, there is a lack of time and resources to change established procedures.

Furthermore, it is not easy to convince data owners of the true value of sharing their registry data. In addition to privacy and security issues, data owners fail to realize the incentives underlying the sharing of their data. To overcome this in the future, financing projects should include clear guidelines to mandate the anonymous sharing of data for research purposes. Including these policies would shed new light on the benefits of sharing patient data on rare diseases to a broader community, truly unlocking its potential.

## 3.7 Summary

This work introduces a semantic web-based layer that provides a holistic perspective over the wealth of knowledge stemming from linked patient registries supported by the growing number of research projects.

Our results are significant in at least three major respects: 1) The use of a model agnostic system, which enables the mapping of patient registries' data from any format to a common shared ontology. 2) The creation of an independent system that can be plugged into any existing patient registry without changing it. This enables the extraction of relevant data elements while maintaining patients' data privacy and security. 3) The adoption of Semantic Web technologies to promote better translation, interpretation, federation and discovery of new knowledge acquired from linked patient registry datasets.

Finally, this solution allows distributed queries to a federated system of linked patient registries. As a result, researchers can easily access a broad set of patient registries just as they would access a single system. We believe this is a milestone towards semantically

interoperable knowledge about rare diseases and will bring us one step closer to personalized medicine.

# Chapter 4

# An automated platform to integrate and publish biomedical data

Publishing, analysing or properly accessing the abundant information resulting largely from experimental studies in the biomedical domain are current challenges for the research community. Problems with the extraction of relevant information, redundant data, and lack of associations or provenance are good examples of main concerns. The innovative nanopublication publishing strategy tries to overcome these issues by representing the essential pieces of publishable information on the SW. However, existing methods to create these RDF-based data snippets are based on complex scripting procedures, hindering their use by the community. Therefore, new and automated strategies are needed to explore the evident value of nanopublications and to enable data attribution mechanisms, an important feature for data owners. To solve these challenges, we introduce the second generation of the COEUS open-source application framework (`http://bioinformatics.ua.pt/coeus/`), an automated platform to integrate heterogeneous scientific outcomes into nanopublications[1]. This results in seamless integration to make data accessible and citable at the same time. No additional scripting methods are needed. A validation of a nanopublishing pipeline is described to demonstrate the system's functionalities, integrating and publishing common biomedical achievements into the SW ecosystem.

[1] This chapter is largely based on the paper by P. Sernadela and J. L. Oliveira, "COEUS 2.0: An automated platform to integrate and publish biomedical data as nanopublications" *IET Software*, vol. 11, Dec. 2017.

## 4.1 COEUS

The COEUS framework was designed to support the integration of heterogeneous life science data, providing an intelligent resource combination mechanism. Moreover, the combined information can be ported to the semantic level using a unique and common data model benefiting distributed information access, as explored in Chapter 3.

Due to its proven features, this framework has been used in diverse scientific projects. For instance, in the rare disease domain, it has been applied to integrate distinct and related resources offering a semantic network for rapid information access [66], to support the creation of new SW tools for distributed knowledge access [75], and also to enable federation strategies from dispersed patient registries and biobanks [15]. In the healthcare context, it has been applied to support the storage and access of semantic radiology annotations [29] and also to enable semantic search over Digital Imaging and Communications in Medicine (DICOM) repositories [178]. Additionally, it has been used in the integration of heterogeneous text-mining annotations for further exploration [28].

In this way, these sample use-case scenarios place COEUS as an essential framework for the scientific community to make traditional data accessible through SW standards. Overall, the new update aims to enable data publishing according to modern data citation strategies, emerging as a vital requirement to make data both accessible and citable for further exploration [26].

## 4.2 COEUS 2.0

The SW paradigm introduces multiple technologies and strategies that are suitable to represent real-world relationships in digital information systems, namely in the life sciences. Moreover, SW standards tackle challenges in the most diverse domains, from data heterogeneity to service interoperability [179], enabling knowledge modelling, sharing and integration [180]. The concept of nanopublications illustrates one of the recent strategies to implement machine-readable knowledge assertions. With this standard still in its infancy, the available transition processes are based on custom scripting solutions that transform original data into the nanopublication format. To automate this process, we redesigned COEUS algorithms and interfaces, offering a more generic, easier and customizable data publishing framework. COEUS 2.0 novel contribution is focused on the development of 4 main components: 1) Implementation of nanopublication generation algorithms; 2) Adaptation of existing methods for data integration; 3) Web-based system development to

support both data integration and nanopublication generation processes; 4) Development of a specialized nanopublication store. Next, we describe these major modifications, streamlining the nanopublication generation and publishing process.

## 4.3 Architecture

The COEUS first generation framework was designed to make traditional data formats accessible through SW standards. It provides a variety of connectors to combine data from diverse sources, providing generic and essential data loading mechanisms for the data publishing architecture. In addition, the integration process is carried out through an organized ontology model (e.g. *Protein-Uniprot-P51587*), simplifying data transformation and access, as data are always available through a hierarchical and well-organized model. Likewise, COEUS has services that are currently fundamental to deliver SW applications. Examples of these are the SPARQL endpoint and the Linked Data methods [81]. The SPARQL endpoint works as a query system, in which we can perform complex queries. The Linked Data interface provides easy access to the information through the Web Browser. Both services will also support the query and retrieve mechanism in the final nanopublication solution.

Figure 4.1 shows the overall nanopublishing architecture, illustrating the main steps of the workflow: from generic data to nanopublications. The pipeline starts by combining the input data, creating one or more resources and their data endpoints (Figure 4.1, block 1). Next, the engine will integrate, generate and store the final output (Figure 4.1, block 2). In the first engine phase, the data will be integrated based on advanced ETL features. The second phase will deal with the transformation process to generate each nanopublication record. The last engine phase will store the nanopublications generated into a named graph store following the nanopublication guidelines. Every nanopublication created will be made publicly accessible by several services (Figure 4.1, block 3). The combination of these steps enables a new nanopublishing pipeline where the tasks are automated. By adopting this architecture, the framework is able to make data usable and citable at the same time.

### 4.3.1 Data integration

The implemented pipeline generates nanopublications from the integrated data. To support data integration our pipeline is based on COEUS ontology model. This is organized

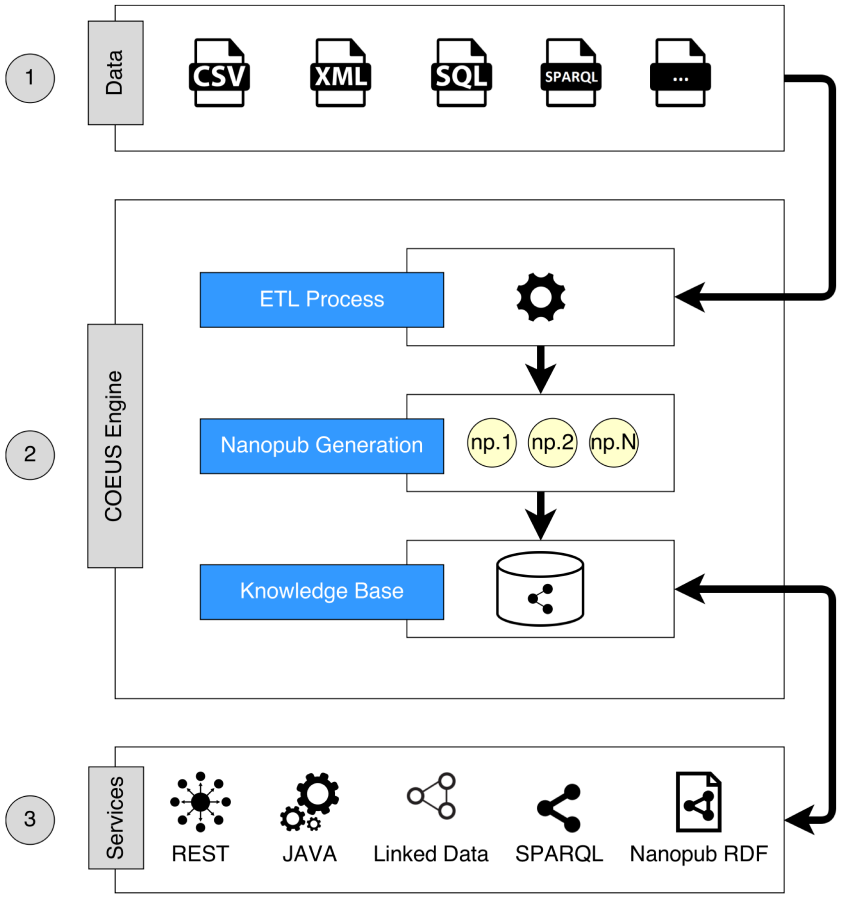


Figure 4.1: Nanopublishing workflow: from generic data to nanopublications. 1) Miscellaneous data studies' formats can be imported. 2) Advanced ETL features integrate information and generate nanopublications. 3) Advanced services are enabled for data query and retrieve.

in a tree-based model: data connections are mapped to *Entity-Concept-Item* structures, which are connected to *Resources*, supporting integration and exploration settings, respectively. To better understand this organization, the ontology model represented in Figure 4.2 and available online at `http://bioinformatics.ua.pt/coeus/ontology/` must be taken into account. This ontology defines *Entities*, *Concepts* and *Resources* used in the data integration process. Additionally, the content is used to support the framework instance configuration, from the management of external resources in the connectors to the labelling rules for each individual *Item*. Essentially, a *Seed* can have several *Entities*, and each *Entity* can be associated with one or more *Concepts*. *Concepts* aggregate exclusive *Items* and are linked to *Resource* information. The data import process uses

*Resources'* properties to load and filter data for the integration layer. This abstraction layer aims to transform the data being integrated into a common model-independent format. In practice, the developed process produces a network for each new *Item*, mapping the configured predicates to the values from the external resources. In this data abstraction layer, the triplification process enables the creation of triple statements from the abstracted data model for further storage in the knowledge base. Further information about the integration process is available in COEUS' website documentation (`http://bioinformatics.ua.pt/coeus/documentation/`).

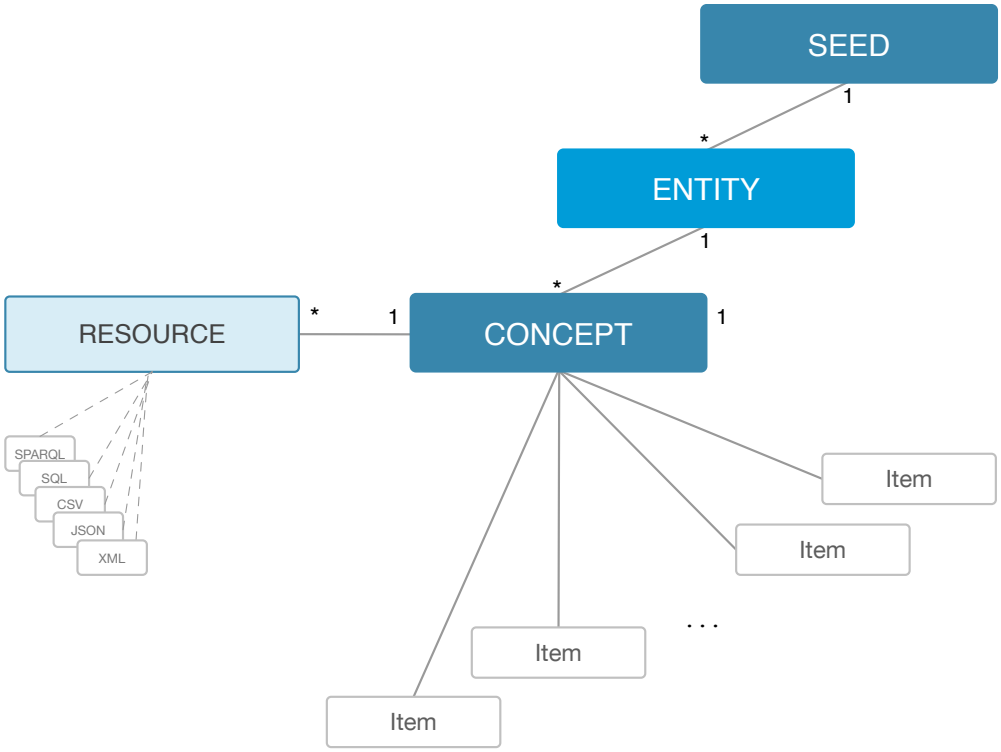


Figure 4.2: COEUS ontology overview.

### 4.3.2 Nanopublication generation

The triplified data are organized in a hierarchical tree structure that the engine must traverse. For this process, the different data and related connections are collected (type of *coeus:concept*) based on the root concept. With this information and associated data (optional), the nanopublications are created.

The COEUS engine adopts the N-Triples (`http://www.w3.org/TR/n-triples/`) structure (*subject-predicate-object*) to build its internal Knowledge Base. According to the nanopublication guidelines, each nanopublication must include an associated context (*subject-predicate-object-context*) – N-Quads (`http://www.w3.org/TR/n-quads/`). For that reason, we extended the COEUS engine to support N-Quads in an independent store: in the generation process we retrieve the triples data from the concept items and associate it with the nanopublication Assertion field (context).

Also, we make use of the same organisation for the Provenance and Publication Info graphs of the Nanopublication (Figure 2.10). To complete the Provenance and Publication Info sections, users have the opportunity to manually insert informative data (using ontologies to describe input information). Additional information is automatically generated, making use of the PROV Namespace (`http://www.w3.org/ns/prov#`) to provide a provenance interchange mechanism in each generated nanopublication. In summary, it includes the system responsible for the assertion's information content (*prov:wasDerivedFrom*), and additional metadata regarding the creation time (*prov:generatedAtTime*) and authorship (*prov:wasAttributedTo*).

When the nanopublication field structure has been adequately generated, the creation pipeline ends with the formation of the nanopublication itself and consequent identifier attribution. Finally, the linkage between the list of nanopublications and their respective concept is completed through the *prov:generated* and *prov:wasGeneratedBy* object properties attribution.

As mentioned, to present data, COEUS has several interoperability features. These include REST services, LinkedData interfaces and a SPARQL endpoint. The formation of a novel nanopublication store required the adoption of a specific approach to retrieve data. In this way, the platform contains a specific exporting format option (represented in Figure 4.1 as "Nanopub RDF"), concordant with the nanopublication schema and accessible through a URI.

## 4.4 Results

In this section, we test the feasibility of COEUS 2.0 as a nanopublishing platform. The presented case study aims to provide the Gene Reference Into Function (GeneRIF) dataset and associated information as nanopublications.

### 4.4.1 Case study

The National Library of Medicine (NLM) started a Gene Indexing initiative in April 2002 with the goal of linking any article about the basic biology of a gene or protein to the corresponding Entrez Gene entry [181]. The result is a growing database entitled GeneRIF (Gene Reference Into Function) within the Entrez Gene database. Each GeneRIF entry is a concise and short phrase (up to 255 characters in length) describing a function related to a specific gene, supported by a publication identifier, the PubMed ID [182]. The publication provides evidence for the assertion text. GeneRIFs can be viewed as "a type of low-compression, single-document, extractive, informative, topic-focussed summary" [142] and public access to this information can be obtained through FTP at `ftp://ftp.ncbi.nih.gov/gene/\gls{generif}/`. Ten thousand GeneRIF-mined entries were used to validate our final solution with the goal of integrating and delivering this dataset as nanopublications. To completely assess our integration pipeline, we defined four key goals:

1. Integrate the CSV GeneRIF dataset (Gene ID, PubMed ID, GeneRIF text, taxonomy ID and last update timestamp) into COEUS knowledge base.
2. Extract additional information from PubMed (publication title, journal title, abstract, etc.) and link this information with the GeneRIF content.
3. Link each PubMed ID to the PubMed website.
4. Generate nanopublications for the collected information.

### 4.4.2 Validation

To provide an easy set-up solution we have developed a web user interface (Figure 4.3), supporting the creation of nanopublications for a broader research community. The employed approach is based on an HTML and Javascript Web Interface (on the client side) that interacts through a REST API layer with the deployed framework (on the server side). The new interface is a vital feature for all stakeholders due to miscellaneous improvements to the installation (e.g. database setup, ontologies used in the mapping process, etc.) and configuration process (e.g. creation of the data integration hierarchical model, ontology property mapping, nanopublication generation, etc.).

Guided by the user interface, two tasks need to be performed to accomplish the objectives mentioned: data integration (Figure 4.4, block 1) and nanopublication

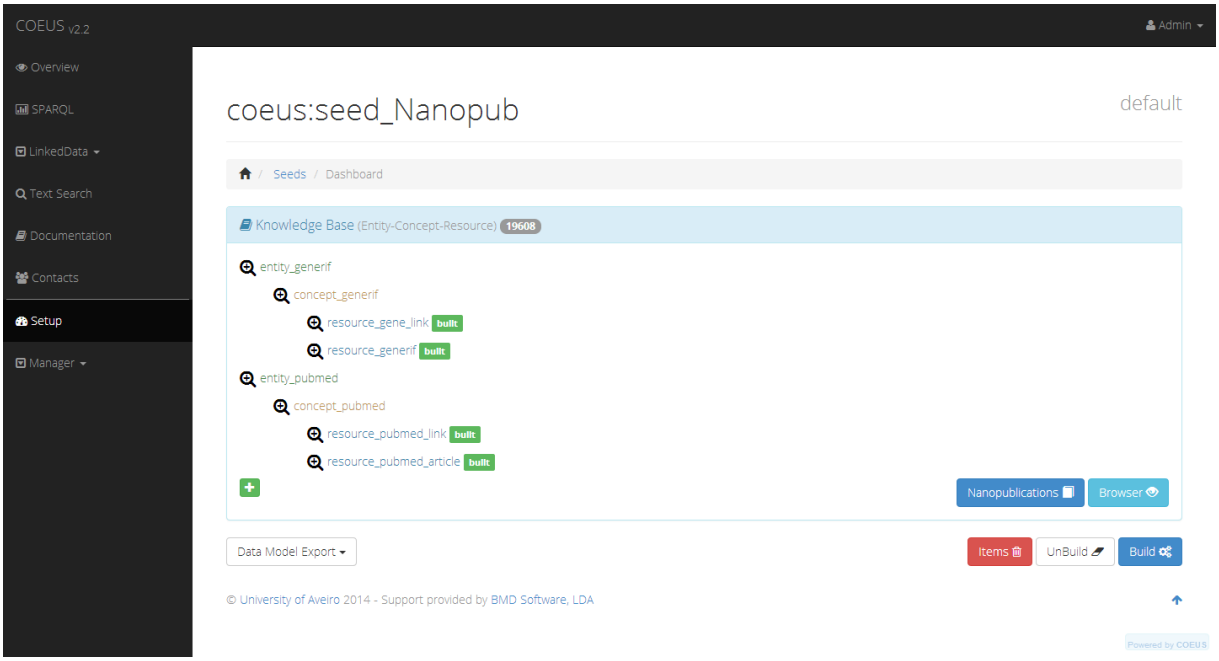


Figure 4.3: Platform web interface overview. This provides a more agile setup for new COEUS installations. Users can now explore the framework without relevant know-how regarding ontologies' creation and management.

generation (Figure 4.4, block 2). A complete user tutorial description of these tasks, i.e. step-by-step guide, is available at `http://bioinformatics.ua.pt/coeus/assets/files/nanopub_tutorial.pdf`.

In the integration task, data are extracted from the GeneRIF dataset, and through an advanced "triplification" process, generated knowledge is stored in a triple store. Next, a combination procedure is applied to integrate additional information from the PubMed website. The result of this integration is the availability of well-structured data according to the COEUS ontology. The integration phase converted and stored the ten thousand GeneRIFs entries and associated information in the KB, generating about three hundred thousand triples.

The second task was nanopublication generation. This stage starts with the selection of data *Concepts* (the GeneRIF *Concept* is selected as *Concept* root) in the nanopublication web interface. The selected data will be mapped to the *np:Assertion* field automatically. For the *np:Provenance* and *np:PublicationInfo* field the information can be added manually. A list of object properties will be suggested when the user starts typing to make the semantic conversion. At this stage, the user can include information about additional

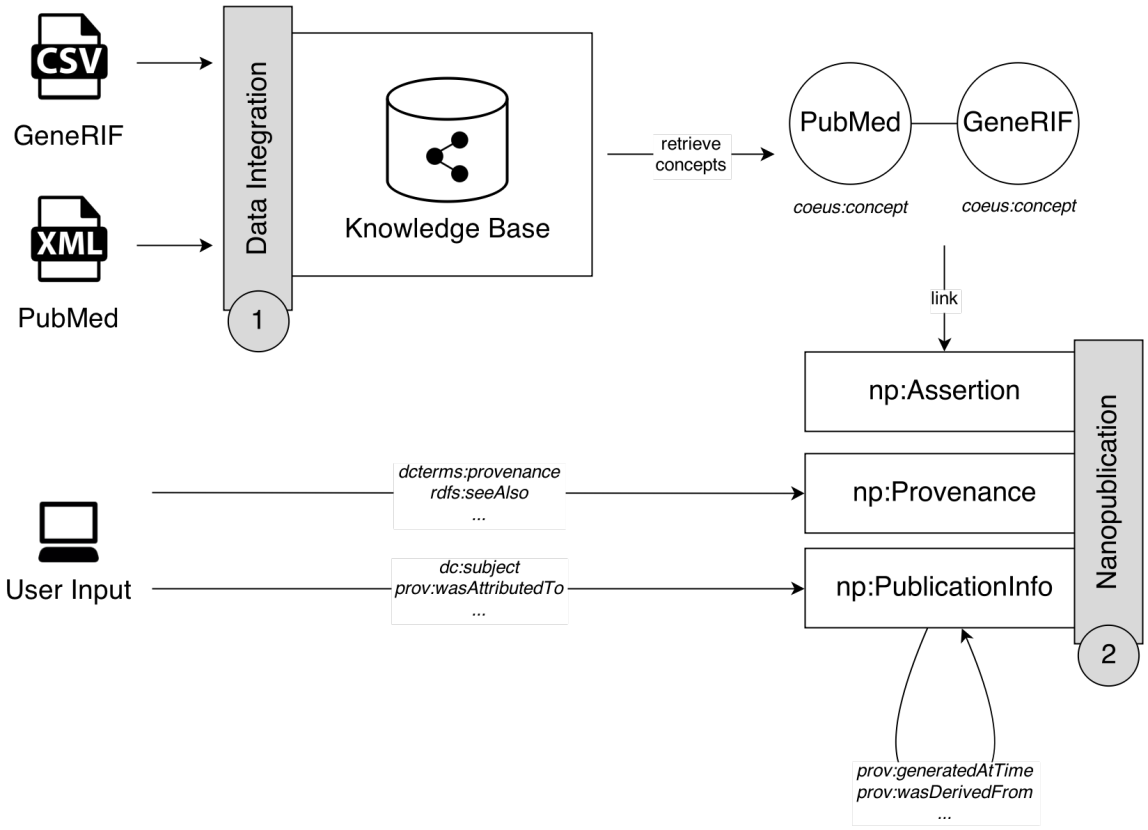


Figure 4.4: Nanopublication generation from GeneRIFs. 1) From the GeneRIF dataset, data are integrated along with PubMed information. 2) The aggregated data are mapped to the assertion field to form the nanopublication corpus. Likewise, information about provenance or additional metadata can be added to further enrich the nanopublications.

provenance and identification such as research ID or email address. However, as mentioned, some properties are automatically included, such as *prov:generatedAtTime* and *prov:wasDerivedFrom*. The nanopublication generation process resulted in ten thousand GeneRIF nanopublications, as expected.

After the two tasks, data are available for query and retrieval using the framework services. Additionally, a nanopublication-specific service is enabled for each nanopublication at "/nanopub/id". This service is accessible via its URI and allows exporting in N-Quads and TriG format, compliant with the nanopublication schema.

The following SPARQL query establishes the link between GeneRIF text and the PubMed article title and website link, included in the nanopublication:

```
PREFIX coeus: <http://bioinformatics.ua.pt/coeus/resource/>
PREFIX np: <http://www.nanopub.org/nschema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX prov:  <http://www.w3.org/ns/prov#>

SELECT ?generif_text ?pubmed_title ?pubmed_link ?nanopub
    ?prov ?derivedFrom{

    GRAPH ?g {
        ?nanopub np:hasAssertion ?assertion .
        ?nanopub np:hasProvenance ?provenance .
    } .

    GRAPH ?assertion {
        ?pubmed_item \gls{rdf}s:seeAlso ?pubmed_link .
        ?pubmed_item dc:title ?pubmed_title .
        ?generif_item dc:description ?generif_text .
    } .

    GRAPH ?provenance {
        ?assertion dcterms:provenance ?prov .
        ?assertion prov:wasDerivedFrom ?derivedFrom
    } .
}
LIMIT 100
```

The output, available for testing at `http://bioinformatics.ua.pt/coeus/sparqler/` shows where the data came from and who was responsible for generating the nanopublication. Retrieving the provenance of the assertion is important to evaluate its trustworthiness.

Moreover, the gene identifier contained in each GeneRIF nanopublication can be mapped to the associated PubMed article. The next SPARQL query describes such interaction associating the gene and the respective article title, i.e. getting the original source for the assertion:

```
PREFIX coeus: <http://bioinformatics.ua.pt/coeus/resource/>
PREFIX np: <http://www.nanopub.org/nschema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?gene ?pubmed_title ?nanopub {

    GRAPH ?g {
        ?nanopub np:hasAssertion ?assertion .
    } .

    GRAPH ?assertion {
        ?pubmed_item dc:title ?pubmed_title .
        ?generif_item dc:relation ?gene .
        ?gene coeus:isAssociatedTo ?generif_item
    }
}
LIMIT 100
```

## 4.5 Discussion

The interactive platform described in this paper aims to improve on currently available hand-scripting tools to create and automatically publish RDF content consistent with the nanopublication standard. This is achieved by following a combination of semi-automated ETL processes where each step is now faster and less error-prone, allowing the transformation of heterogeneous data sources. The generation of each nanopublication follows an automated strategy that requires no manual modelling of the *assertions*. To model the *Provenance* and *Publication Info* graphs, our extension includes customizable interfaces that can be completed with user-desired ontologies. Implemented features are not available in any state-of-the-art tool, which are usually based on scripting solutions for each different problem. Our solution substantially augments the ability of non-informatician researchers to produce nanopublications from their data studies, while maintaining the respective credit. At same time, they can promote data-sharing standards using adequate mechanisms such as SPARQL and Linked Data interfaces, making their nanopublications discoverable and accessible through multiple methods. Regarding these features, we believe that our second version framework will improve current publishing methodologies making data both citable and accessible through modern standards.

## 4.6 Summary

The great evolution of scientific data produced year by year, including experimental data, begs for new approaches to grasp novel scientific results. The nanopublication standard is the SW solution to this issue, allowing the summarization and linkage of scientific outcomes, while ensuring the appropriate data attribution. The rise of this prominent standard quickly triggered the need for new data transformation tools. However, most of the available solutions are prototype and scripting solutions, each one targeting a specific domain. In contrast, the second generation of COEUS introduces a novel automation level, enabling the generation of nanopublications from generic data sources. This new version makes the modelling effort feasible for researchers, reducing errors and encouraging them to publish and integrate their results as nanopublications. Moreover, it includes customizable options that can be combined with external ontologies providing additional mapping to each structured field that composes the nanopublication. Study results, such as GeneRIF, available in common formats, can be easily incorporated into this framework. With our new nanopublishing workflow, users can translate their data into our engine, selecting and mapping the essential structured fields easily. Likewise, it provides an attribution system with proper recognition of the authors, enabling appropriate data-sharing mechanisms, according to Linked Data principles. As such, our renewed platform will benefit the research community and promote data-sharing standards.

# Chapter 5

# Semantic-based architecture for biomedical literature annotation

Knowledge extraction from the biomedical literature plays an important role since most of the relevant information from scientific findings is still maintained in text format. In this endeavour, computational annotation tools can assist in the identification of biomedical concepts and their relationships, providing faster reading and curation processes, with reduced costs. However, the separate usage of distinct annotation systems results in highly heterogeneous data, as it is difficult to efficiently combine and exchange this valuable asset. Moreover, despite the existence of several annotation formats, there is no unified way to integrate miscellaneous annotation outcomes into a reusable, sharable and searchable structure. Taking up this challenge, we present a modular architecture for textual information integration using Semantic Web (SW) features and services[1]. The solution described in this chapter allows the migration of curation data into a common model, providing a suitable transition process in which multiple annotation data can be integrated and enriched, with the possibility of being shared, compared and reused across semantic knowledge bases.

## 5.1 Architecture

In the Chapter 2, we have discussed several alternative methodologies to represent text-mining annotations. Although major contributions have been made in this area, it is

[1] This chapter is largely based on the paper by P. Sernadela and J. L. Oliveira, "A semantic-based workflow for biomedical literature annotation", *Database*, vol. 2017, p. bax088, Jan. 2017.

still challenging to adapt and link the output of these distinct tools. To address this issue, we implemented a modular architecture able to support the integration of annotations from multiple extraction tools into the SW ecosystem (Figure 5.1).

The proposed approach aims to provide a seamless transition from unstructured information to the SW level. The overall architecture is based on a modular and pipelined approach, divided into three interconnected, though independent, components: 1) knowledge discovery; 2) semantic integration; and 3) semantic services.

### 5.1.1 Knowledge discovery

In this component, textual documents are examined using state-of-the-art text-mining methods for the identification of relevant concepts, respective attributes and relationships. These extraction techniques can be applied by one or a combination of automated text-mining tools. This means that the architecture does not rely on a single text-mining solution to perform information extraction, with it being possible to aggregate results from several systems. However, each text-mining solution must be delivered as a RESTful Web service to be compliant with the implemented architecture. The deployment of those resources through Representational State Transfer (REST) services allows us to standardize how HTTP requests can be performed within the architecture. Service invocations are made through HTTP POST requests, accepting *text/plain* as content type. This simplifies communication between the components developed and facilitates the configuration process for additional text-mining tools and systems integration. The implemented architecture supports NER systems, complete concept recognition systems, and relation extraction systems. In the section 5.2, a setting with two distinct text-mining solutions is assessed,

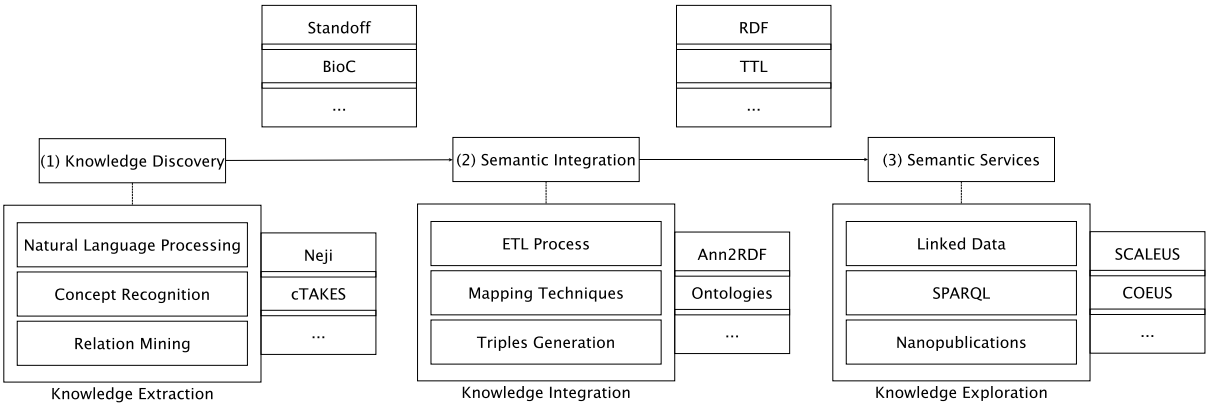


Figure 5.1: Semantic-based architecture for scientific information integration.

dealing with different formats and results.

### 5.1.2 Semantic integration

Information extraction tools produce several annotation formats. The migration of this data into SW format and services provides additional value regarding the share of knowledge. To allow this transition, we have developed the Ann2RDF system [31]. Ann2RDF (`http://bioinformatics-ua.github.io/ann2rdf/`) is based on the creation of modular integration algorithms to deal with the different formats resulting from text-mining tools. The ability to acquire data from several and miscellaneous annotation formats benefits developers, allowing each one to implement and integrate their format in a common interface. Developed algorithms are based on Object Relation Mapping (ORM) techniques for mapping different data structures to a single representation and on ETL procedures to select and extract annotations' content based on regular expressions and data parsers such as XML Path Language (XPath). Currently, the system supports the integration of most BioNLP (`http://bionlp.org`) formats out-of-the-box such as the BioC and Standoff formats, with it also being possible to additionally customize new formats.

After this selection and extraction processes, annotations' objects are semantically enriched by using ontology mapping procedures: the system makes use of an external JSON-based configuration file to assist the ontology mapping process. In this configuration file, the mappings between classified concept categories and relation properties (i.e. associations between concepts) are defined to the respective ontology terms. This allows standardization of annotations' content, e.g. "*A relatedWith B*" to "*A dc:relation B*", using for instance, the Dublin Core Ontology [183].

Next, there is the possibility of normalizing the detected concepts. Due to the existence of many NER tools that do not include concept normalization tasks, the system offers an optional normalization service. The invocation is also performed in the same configuration file, declaring external HTTP POST requests. For this invocation, two properties are needed: the service location and the regular expression to apply to select the desired output. With this external support, services such as BioPortal Annotator [184] (e.g. *service*: "`http://data.bioontology.org/annotator?apikey=XXXX`", *query*: "[*].annotatedClass.@id" ) or BeCAS [179] (e.g. *service*: "`http://bioinformatics.ua.pt/becas/api/text/annotate`", *query*: "*.*.refs") can be easily integrated, providing an enhanced incorporation of the annotated data and improved simplification for the semantic integration process.

Finally, harmonization methods are responsible for performing an adequate linkage between extracted content and the respective structured model. To represent the processed data, our architecture model is based on Annotation Ontology (AO) [180], an open representation model for representing interoperable annotations in RDF which is currently being used by the W3C community (`https://www.w3.org/TR/annotation-vocab/`). It provides a robust set of methods for connecting web resources, for instance, textual information in scientific publications, to ontological elements, with full representation of annotation provenance, contextual metadata describing the origin or source [181, 182]. By linking new scientific content to computationally defined terms and entity descriptors, AO helps to establish semantic interoperability across the biomedical field. Through this model, existing domain ontologies and vocabularies can be used, creating rich stores of metadata on web resources.

### Concept Model

We reuse the AO core ontology components to describe generated annotations. In Figure 5.2, we present the adopted core model, using a sample annotation regarding identification of Alzheimer's disease. The central point of the representation includes the URI (e.g. *ann2rdf:T1*), the document source (e.g. Pubmed ID *25766617*), and the respective annotated data (e.g. *Alzheimer Disease*). The text selectors are used to identify the string detected on the document: the *ao:exact* data property represents the linear sequence of characters, i.e. the subject of the annotation, the *ao:offset* data property indicates the distance from the beginning of the document up to a given element or position, and the *ao:range* data property represents the number of characters starting from the offset. Information about the annotation itself is connected through two different properties: the *ao:body* representing the annotated resource and the *ao:hasTopic* indicating the semantic identifier of the detected resource (e.g. OMIM ID *104300*). The identifier is attributed by the normalization service to represent *"Alzheimer Disease"* annotation due to the inexistence of such information on the previously annotated data. If the annotation data already contemplate a semantic identifier, it is extracted and connected to the annotation graph. Moreover, the annotations are linked to the respective document source through the object property *ao:onSourceDocument* providing a provenance interchange mechanism. By using this simplified model, entity annotations can be easily mapped to a SW-compliant format.

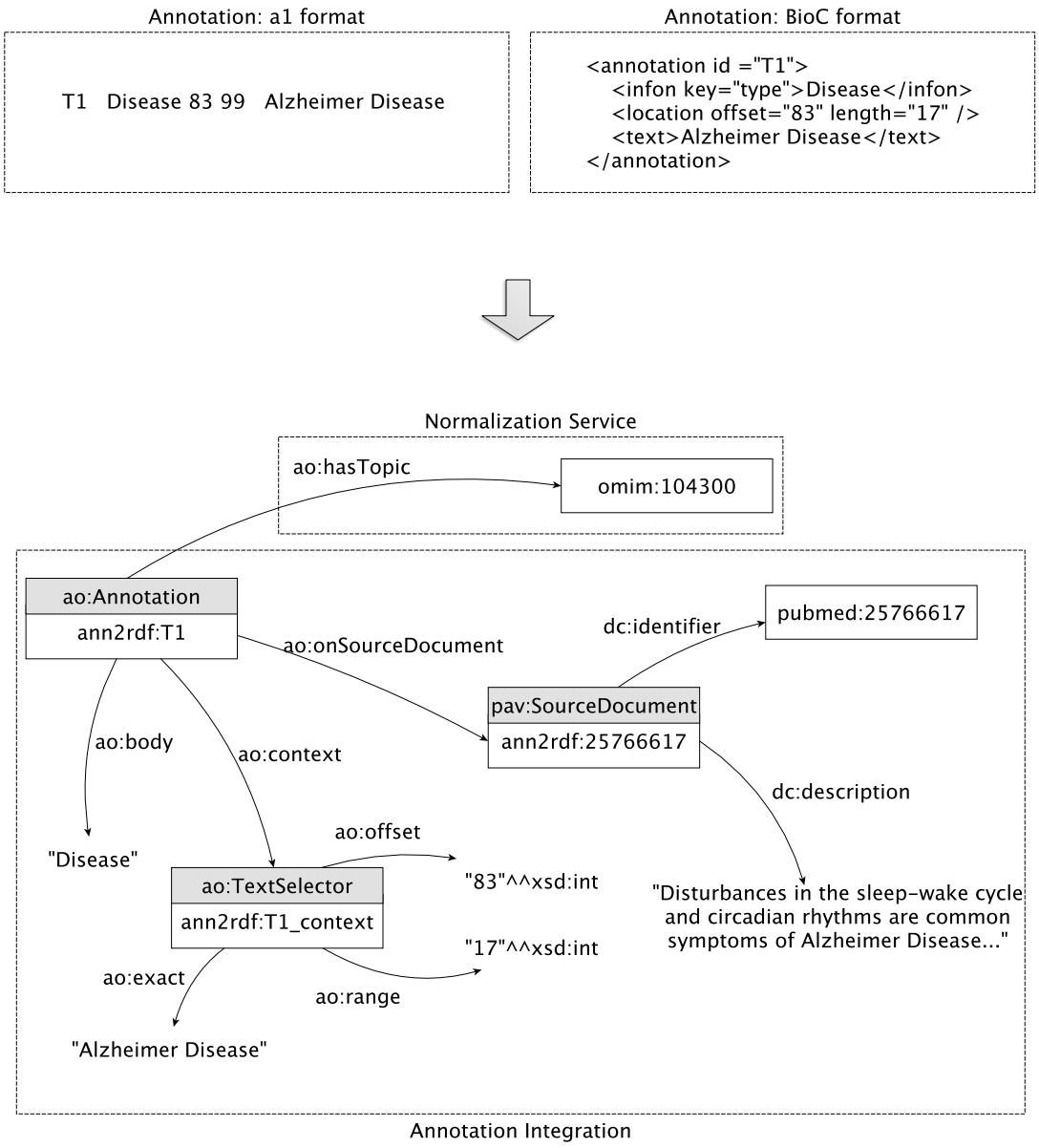


Figure 5.2: Annotation model: sample extraction of the integration and representation of an annotation related to "Alzheimer disease" using two distinct data formats.

#### Relation Model

Researchers typically refer to relation extraction as the task of identifying binary relations between concepts, and to event extraction as the identification of more complex relationships, involving verbs or normalized verbs (i.e. trigger) to characterize the event type. Event extraction techniques started to become more common with the introduction of BioNLP shared tasks [142], allowing the construction of complex conceptual networks.

We introduced new relationships to allow the representation of annotation interactions. To represent the relations (Figure 5.3), our model essentially connects the binary entities through one additional annotation. The relation is not directly established between the two entities involved due to the possible existence of different specificity in the object property linkage between relations. For this reason, a new annotation is created to associate the two annotations and a respective descriptive relation type is attributed through the *ao:body* property. Regarding the representation of events, our model achieves a similar structure of the relation annotations but with some adjustments, i.e. instead of only representing the binary relation it can represent multiple associations between annotations. Using the representations described, the outcomes of text-mining tools can be easily integrated into a unified model providing SW interoperability features for the mined resources.

### 5.1.3 Semantic services

The SW has gained an increasing role as a suitable environment to solve knowledge representation and interoperability problems, creating accessible and shareable information across application and database boundaries. Its adoption by the life science community allows better standards and technologies to be delivered, making the interconnection across knowledge domains possible and effective. Taking those benefits into account, our flexible solution enables the deployment of several semantic-based systems and services. Developed to support the current need of semantic-web services [147], existing systems explore the potential behind SW technology, enabling the quick creation of new knowledge bases for further exploration. COEUS [20], Scaleus [32] and SADI [108] are some examples of these systems, which can be used along our modular solution.

However, this work is only focused on the implementation and exploration of services residing in the Scaleus web system (described in Chapter 6). With this adoption, we take advantage of several services, including a database management system with simplified APIs, a SPARQL query engine supporting real-time inference mechanisms, and optimized

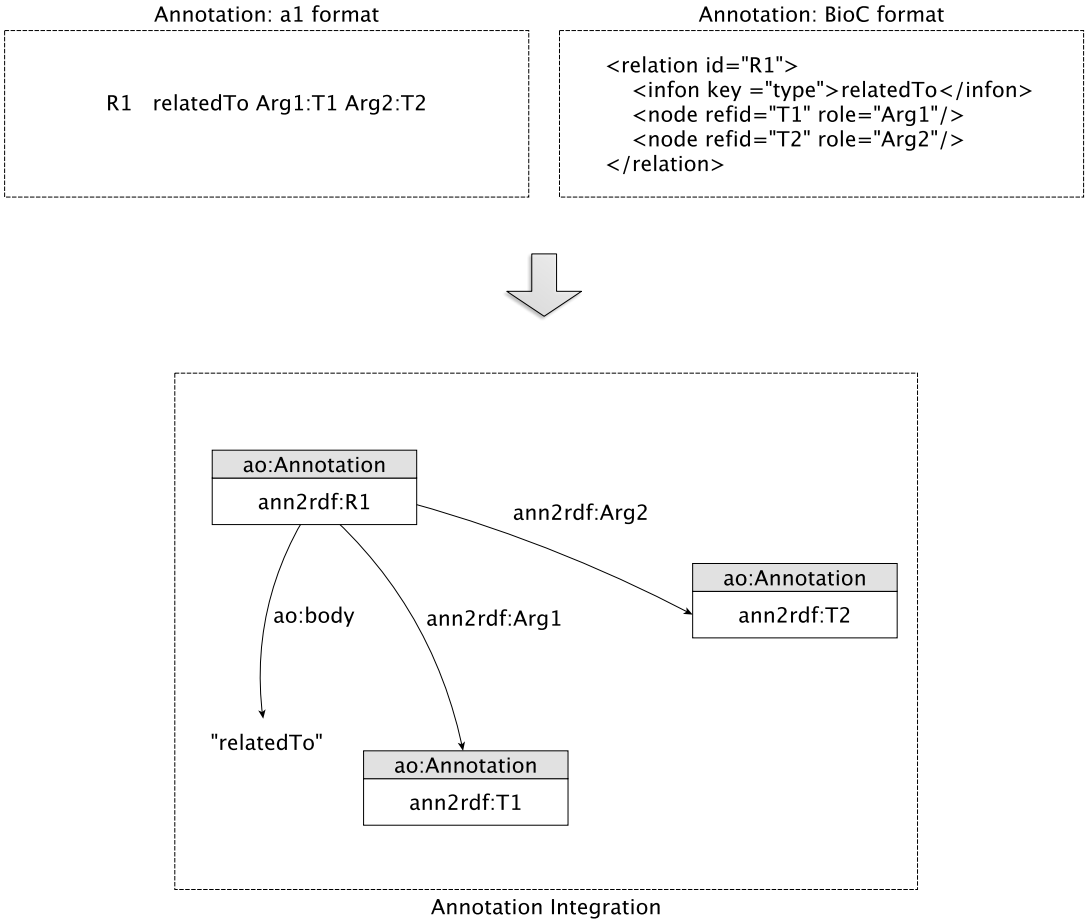


Figure 5.3: Relation model: sample extraction of the integration and representation of a *relatedTo* annotation relationship using two distinct data formats.

text searches over the knowledge base. Inference on the SW is one of the most useful tools to enhance data integration quality, automatically analyzing the content of the data and discovering new relationships. In the deployed system, the SPARQL query engine plus user-defined rules makes it possible to generate new relationships from existing triples, and therefore increase reasoning capabilities by inferring or discovering additional facts about the stored data. Regarding the text-search feature, it offers the ability to perform free-text searches within SPARQL queries. The support of SPARQL Federated Query (`https://www.w3.org/TR/sparql11-federated-query/`) is also an available feature allowing the execution of distributed queries over different SPARQL endpoints. In this way, the deployment of these semantic services with the combination of existing life science knowledge bases such as the Bio2RDF [13] or the EMBL-EBI RDF Platform [14] provides

a well-structured network, in which federated inquiring mechanisms can be easily applied [15, 16]. Scaleus detailed features will be described in chapter 6.

## 5.2 Results

The developed architecture involves a diverse combination of systems and technologies, lying in the intersection of knowledge discovery and SW methods. Due to its modularity, several components can be used, providing greater freedom for end-users and offering distinct possibilities for information integration and access.

Regarding the contribution, this work is focused on the implementation of a modular semantic-web workflow for the integration and reuse of multiple text-mined results. To allow this, three main components were developed: 1) Development of literature extraction methods based on RESTfull APIs; 2) Improvement and adaptation of Ann2RDF algorithms for annotations' integration and enrichment. 3) Development and deployment of a Scaleus instance, for annotations' exploration (available at `http://bioinformatics.ua.pt/dmd/scaleus/`). In the next sections, we explore and evaluate these components towards a unified workflow for data integration and distribution.

### 5.2.1 Information extraction

To demonstrate the feasibility of the implemented solution, we explored a combination of two distinct text-mining solutions.

The first solution is Neji [8], a modular framework for biomedical natural language processing. This open-source framework allows the integration in a single pipeline, as dynamic plugins, of several state-of-the-art methods for biomedical natural language processing, such as sentence splitting, tokenization, lemmatization, part-of-speech, chunking and dependency parsing. The concept recognition tasks can be performed using dictionary matching and machine learning techniques with normalization. This framework implements a very flexible and efficient concept tree, where the recognized concepts are stored, supporting nested and intersected concepts with one or more identifiers. The architecture of Neji allows users to configure the processing of documents according to their specific objectives, providing very rich and complete information about concepts.

The second tool used in this example is cTAKES [185], an open-source NLP system for information extraction from free text of electronic medical records. The system was designed to semantically extract information to support heterogeneous clinical research.

It consists of a sequence of modular components (including sentence boundary detector, tokenizer, normalizer, part-of-speech tagger, shallow parser and named entity recognition) that process clinical free-text, contributing to a cumulative annotation dataset. cTAKES was already optimized to explore the characteristics of clinical narratives. By exploring both tools, we expect to maximize coverage in the biomedical and healthcare fields.

Neji and cTAKES services were both deployed with end-user web interfaces and REST APIs, simplifying the test and validation of our architecture (Figure 5.4). The dictionaries used in both solutions were retrieved from the 2014 version of UMLS Metathesaurus database [186], which contains key terminology, classification and coding standards assigned to terms. Each term has a Concept Unique Identifier (CUI), to be assigned to each identified concept. Both solutions can perform concept recognition through REST services.

Additionally, the cTAKES annotator can execute relation extraction techniques between identified concepts. These binary relations are recognized using a rule-based and machine learning components, making it possible to detect interactions such as the degree of (e.g. degree of pain) or location of (e.g. location of pain).

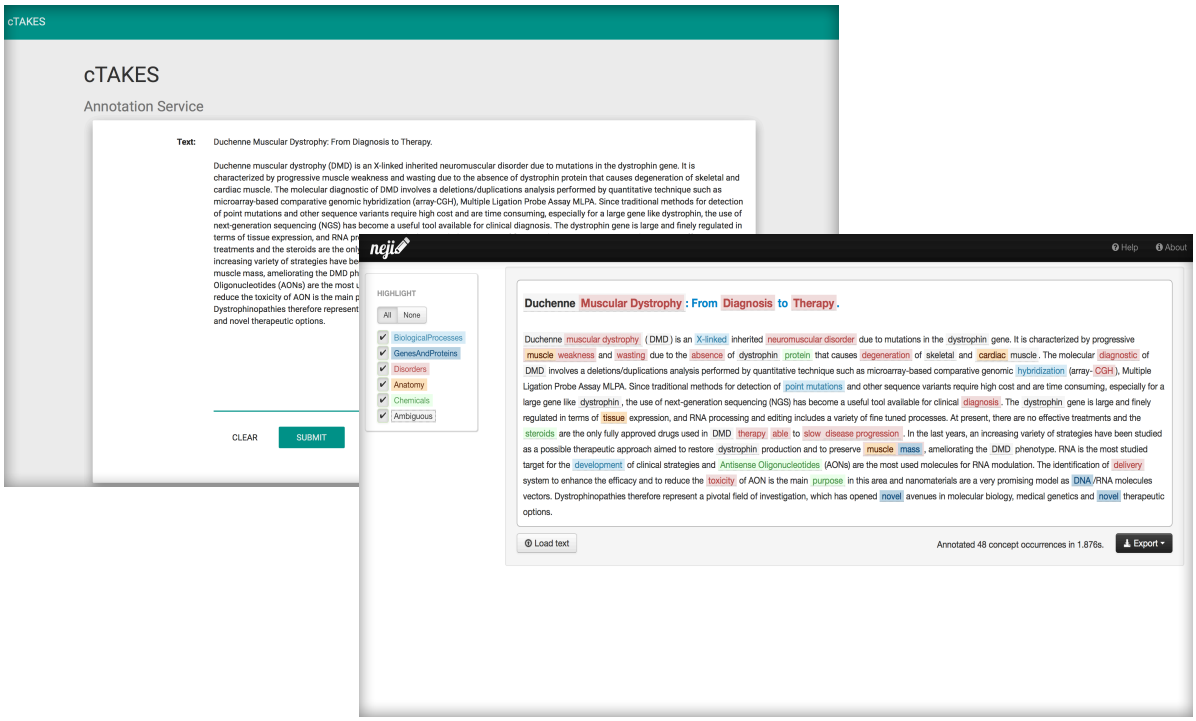


Figure 5.4: cTAKES and Neji developed web interfaces.

### 5.2.2 Evaluation

To validate our architecture, we conducted a case study aimed to create a semantic repository from a dataset related to Duchenne Muscular Dystrophy (DMD), a rare disease condition affecting 1 in 5000 males at birth. For this case study, we collected a dataset containing 2783 DMD related abstracts, obtained by accessing the Entrez Programming Utilities interfaces in the NCBI database.

Figure 5.5 shows our modular workflow. The workflow demonstrates that we take advantage of several annotation tools to extract concepts and relations from the textual information. In this case, the cTAKES delivers respective annotations in the standoff format (`http://2013.bionlp-st.org/file-formats`), where the annotations are stored separately from the annotated text, and the Neji system supplies annotations in the BioC format, a verbose XML format for data exchange.

Using Ann2RDF [31], all the resulting annotations can be integrated into a common and sharable interface. Concepts and relations are independently extracted from the annotation data through advanced ETL processes. Ontology mapping procedures can also be used to enrich the integrated data through configuration properties - annotation tag mappings

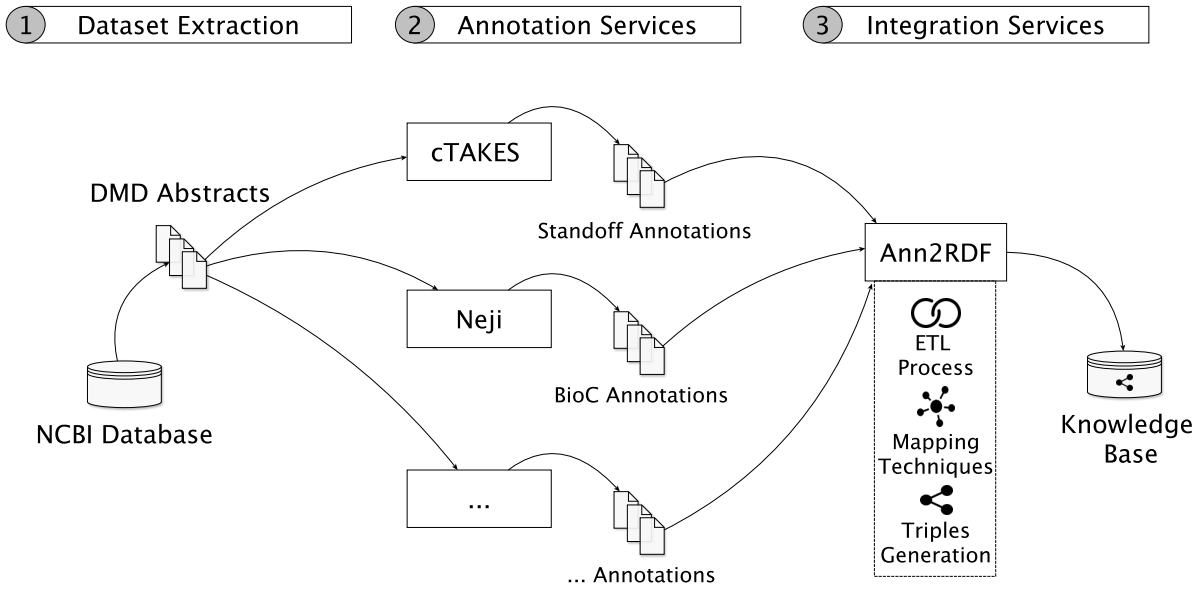


Figure 5.5: Validation workflow overview. 1) Dataset is extracted from the NCBI database. 2) Neji and cTAKES API services were used for information extraction, generating diverse outputs and formats. Additional annotation services can be used. 3) Annotations are forwarded and integrated into a unified model and stored in an accessible knowledge base.

(i.e. classified concept categories, not concept semantic identifier) and property mappings (i.e. associations between concepts) are supported. For instance, if an entity term is recognized as a *Gene_expression* tag, the system allows this linkage to be enriched by adding new mappings to terms available in an adequate ontology (e.g. Gene Regulation Ontology - `http://purl.bioontology.org/ontology/GRO#GeneExpression`). Moreover, it is possible to configure external services to enrich the detected entities with normalization and disambiguation features.

These integration mechanisms are responsible for performing an adequate linkage between the information extracted by the text-mining tools and the respective adopted model. The entire workflow generated a unified knowledge base with more than 3.5 million triples of concepts, relations and respective provenance information (Figure 5.6).

Finally, the integrated information can be combined with existing and related knowledge due to its compatibility with SW standards and queried over SPARQL engines. For instance, it is very straightforward to find the documents where a specific concept was identified (e.g. Skeletal muscle atrophy):

```
PREFIX ao: <http://purl.org/ao/>
PREFIX umls: <http://linkedlifedata.com/resource/umls/id/>

SELECT DISTINCT ?source {
    ?annotation a ao:Annotation .
    ?annotation ao:hasTopic umls:C0234958 .
    ?annotation ao:onSourceDocument ?source .
}
```

The knowledge base from this example can be explored through a set of semantic services available at (`http://bioinformatics.ua.pt/dmd/scaleus/`). Access is provided through a SCALEUS [32] instance, offering a public SPARQL endpoint with data federation capabilities and supporting real-time inference mechanisms. Optimized text searches over the knowledge base are also available.

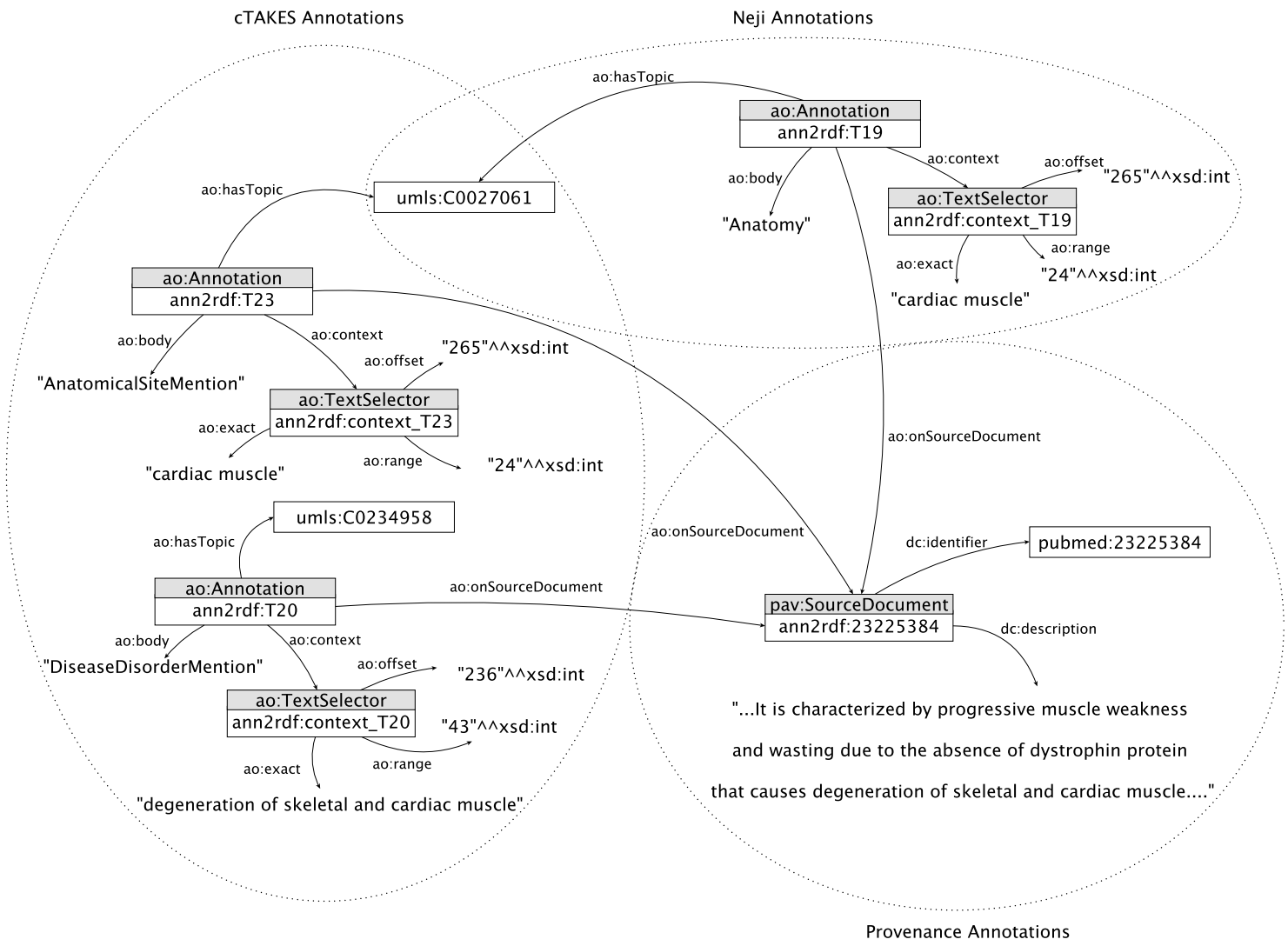


Figure 5.6: Knowledge base sample annotation model. The annotators involved share concept attributions (i.e. *umls:C0027061*), increasing the likelihood of being correctly identified.

## 5.3 Discussion

In recent years, the number of biomedical information extraction systems has been growing steadily. The latest approaches use computational tools to help in the extraction and storage of relevant concepts, as well as their respective attributes and relationships. The product of these complex workflows provides valuable insights into the overwhelming amount of biomedical information being produced. However, interoperability issues in this domain are critical. In this manuscript, we propose an architecture to unify document curation results and enable their proper exploration through multiple interfaces geared towards bioinformatics developers and general life science researchers. This enables a unique scenario where heterogeneous results from annotation tools are harmonized and further integrated into rich semantic knowledge bases.

Compared to existing techniques, our approach integrates several main features:

- The possibility to use and combine text-mined information from different and independent annotation tools.
- The adoption of a unique and effective ontology model that is currently being used by the W3C community.
- The provision of enriched information resulting from the ontological terms mapping process and the combination of text-mined results.
- Fast creation of semantic-powered knowledge bases.
- Information sharing mechanisms are simplified by using SW standards and adequate provenance methods.
- Finally, it enables the exploration of a multitude of SW technologies and services such as reasoning capabilities, Linked Data and SPARQL query endpoints.

Taking advantage of these features, we have implemented a case study regarding Duchenne Muscular Dystrophy (DMD) disease, resulting in the integration of results from two text-mined solutions. The outcome is a fully-connected knowledge base of annotations allowing the exploration of complex interactions between the identified concepts. Additional semantic services combination empowers our final results, delivering enhanced information sharing and discovery methods.

Ultimately, the approach developed envisages providing a modular architecture for textual information integration, normalizing access and exploration. Moreover,

the possibility to combine information from several annotation tools allows enhanced forthcoming quality controls, resulting in a fast strategy to identify gaps between the mined information. Using this approach, information can be easily compared, differentiated and measured according to the user's needs.

Finally, the general architecture of the solution allows its application in the most diverse life science scenarios. For instance, our approach was also used to convert 16 thousand textual radiology reports into a knowledge base with more than 6.5 million triples [29]. In that case, narrative reports were extracted from a SQL database and processed with just one text-mined solution. The outcome was a radiology knowledge base of clinical annotations, currently being used for medical decision support purposes.

## 5.4 Summary

Information extraction systems have been increasingly adopted to facilitate the processing of textual information. The heterogeneity of these tasks, regarding the extraction process, generates a vast quantity of miscellaneous data, which are dependent on the systems used and, in most cases, are not interoperable.

Despite current research efforts, advanced exploration, integration or comparison of these valuable data have been left outside the research path. We proposed a modular framework where these limitations can be overcome. Our solution resides in a fast mechanism to integrate knowledge extracted from several text-mining solutions, enabling the easy creation of semantic-powered databases. The ability to process annotations from several, miscellaneous annotation formats benefits accessibility methods, allowing the integration of heterogeneous formats into a common and interoperable model. This is the major outcome of the implemented solution.

To validate our system, we extracted annotations from the scientific literature, using two different text-mining solutions, leading to the creation of a unified semantic knowledge base. Data exploration methods can be easily applied through several services, making the analysis of extracted knowledge feasible. The repository created follows Linked Data standards, facilitating the application of modern knowledge discovery mechanisms (e.g. reasoning).

# Chapter 6

# Semantic Web services integration for biomedical applications

Recent years have witnessed an explosion of biological data resulting largely from the demands of life science research. The vast majority of these data are freely available via diverse bioinformatics platforms, including relational databases and conventional keyword search applications. This type of approach has achieved great results in the last few years, but proved to be unfeasible when information needs to be combined or shared among different and scattered sources. In recent years, many of these data distribution challenges have been solved with the adoption of the semantic web. Despite the evident benefits of this technology, its adoption introduced new challenges related to the migration process, from existent systems to the semantic level. To facilitate this transition, we have developed Scaleus, a Semantic Web migration tool that can be deployed on top of traditional systems in order to bring knowledge, inference rules and query federation to the existent data[1]. Targeted at the biomedical domain, this web-based platform offers, in a single package, straightforward data integration and semantic web services that help developers and researchers in the process of creating new semantically enhanced information systems. Scaleus is available as open source at `http://bioinformatics-ua.github.io/scaleus/`.

[1] This chapter is largely based on the paper by P. Sernadela, L. González-Castro, and J. L. Oliveira, "SCALEUS: Semantic Web Services Integration for Biomedical Applications", *Journal of Medical Systems*, vol. 41, no. 4, p. 54, Apr. 2017.

## 6.1 Architecture

Scaleus is a semantic web migration tool designed to be simple to deploy and use. The solution is tailored to help users in the creation of new semantic web applications from scratch. Through it, users have access to a set of semantic services for data integration, management and respective exposure according to the Linked Data principles [81]. In a single package, we include a triplestore supporting multiple independent datasets, simplified API and services for data integration and management, and a SPARQL query engine, supporting real-time inference mechanisms and optimized text searches over the knowledge base. The solution was built focused on the development of vital components regarding semantic web application deployment. These components are divided into three main layers: knowledge base, abstraction and services (Figure 6.1). The knowledge base layer is powered by a Transactional Database (TDB) for RDF storage. This provides a high performance and transactional triplestore capable of handling efficient queries by means of a native store. When accessed using transactions, data are protected against corruption, unexpected process terminations and system crashes. The storage mechanism also supports multiple independent datasets providing a structural way to organize several graphs in a single system application. Each dataset backed by the TDB is stored in a single directory in the file system. Likewise, an index directory is recorded beside each dataset to allow even more efficient text search queries. Regarding the abstraction layer, we support our system with Apache Jena (`jena.apache.org`), an open-source Java framework for building SW and Linked Data applications. The abstraction layer implements the required methods to manage the semantic datasets, including the provision of several data loading and query mechanisms. Finally, the services layer implements a HTTP REST interface, which provides an easy and scalable boundary across the RDF store and the user web interface (Figure 6.2). The API provides several methods for triples management as well as a standard SPARQL engine for querying the RDF models, supporting federated queries over different endpoints. These features are orchestrated by a Jetty application server in embedded mode for fast and easy deployment, without compromising the system's overall performance.

### 6.1.1 REST API

Scaleus is powered by a JSON-based RESTful API for data management. A complete list of the methods can be found on the website documentation page (available

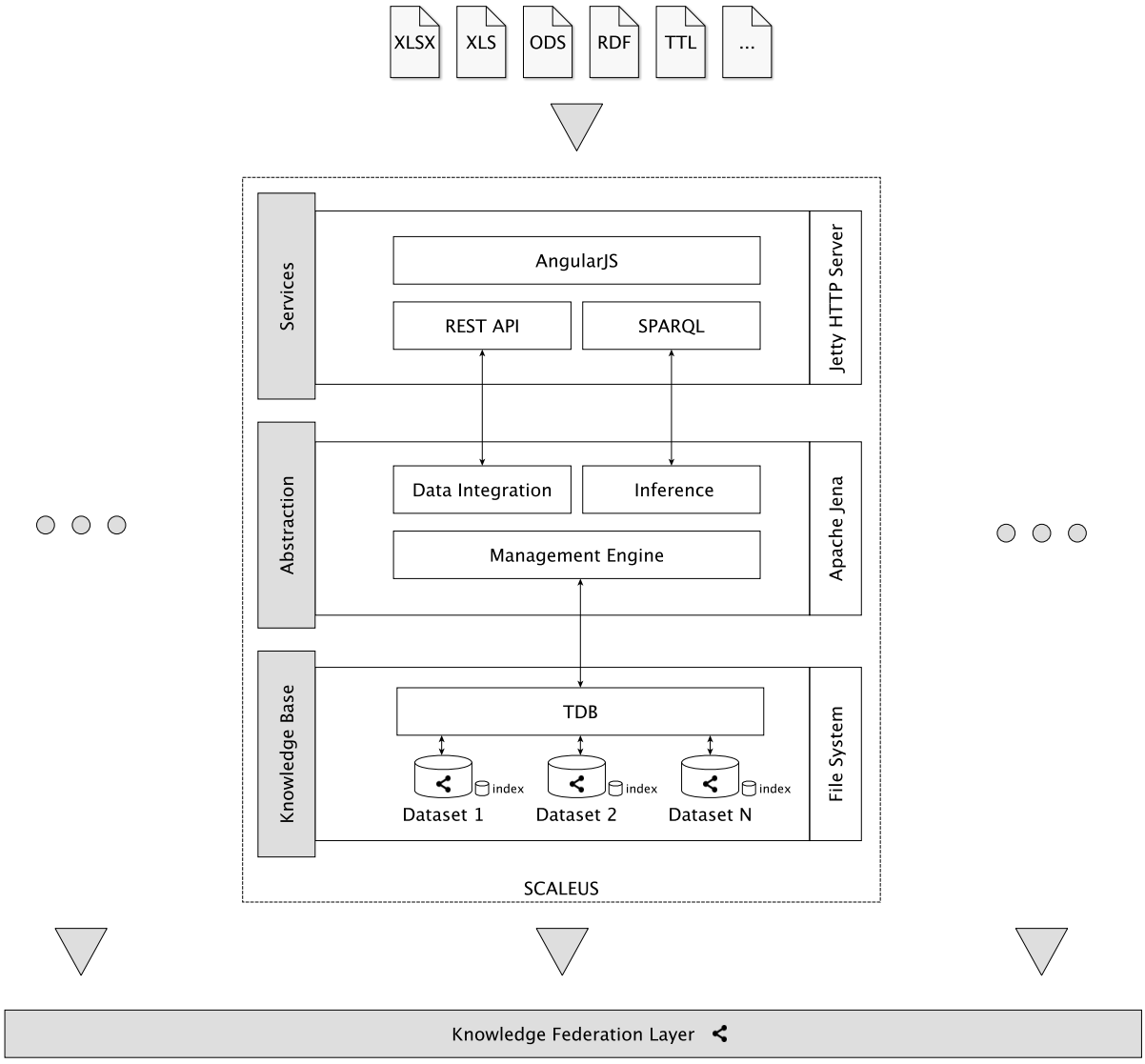


Figure 6.1: Scaleus architecture overview divided into 3 main layers: Knowledge Base, Abstraction and Services. The SPARQL endpoint enables query federation through multiple Scaleus instances.

at `https://github.com/bioinformatics-ua/scaleus#documentation`). Compared to implementation of other existing triplestores, it does not use the W3C standard SPARQL Update Language [187] for graph/triples management, adopting a simpler approach to insert, update or remove RDF triples through the REST API. Despite this simplification, the API maintains full compliance with the SPARQL 1.1 Query Language specification [188].

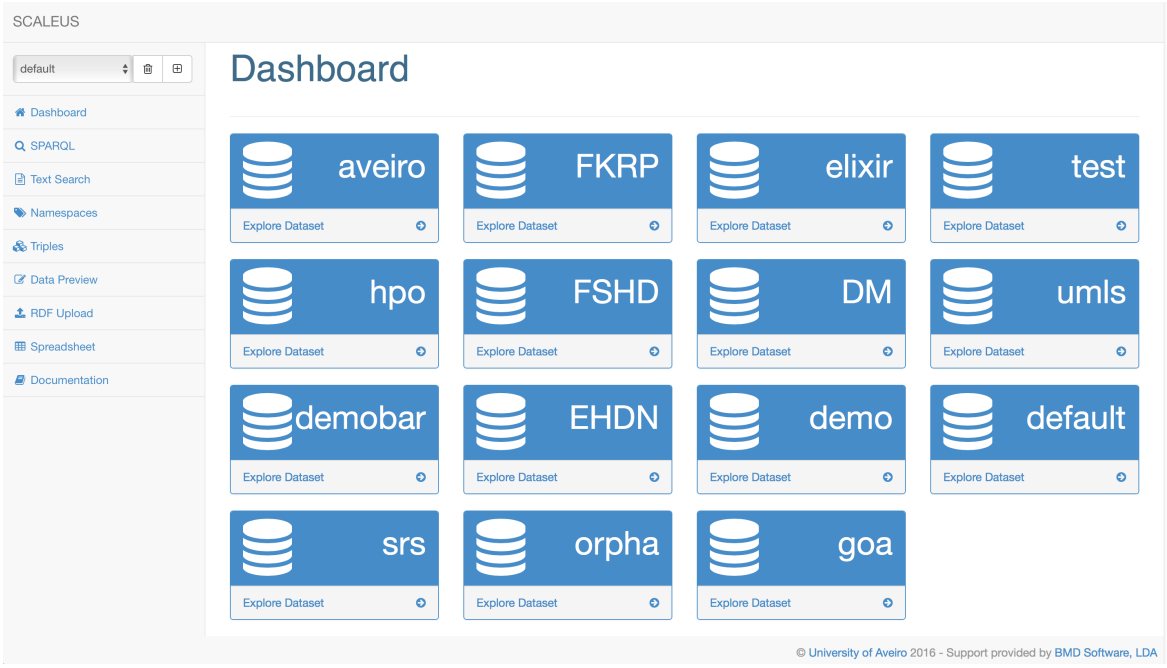


Figure 6.2: Scaleus web interface. The dashboard view shows the complete list of datasets available. Datasets are managed using the left sidebar, including several services for data loading and querying.

### 6.1.2 Data Integration services

Considering the increasing availability of data, there is a clear need for improved user-friendly tools and services targeting the integration of heterogeneous datasets [10]. These tasks are, in most cases, not easy to achieve without the development of complex data integration tools that are unsuitable for adoption by most of the research community. Furthermore, the requirements of Linked Data and associated ontology models makes the transition to a common and sharable structure adaptable to each life science domain even more difficult. With Scaleus, we attempt to simplify these migration tasks by providing a set of data connectors and interfaces that help in the translation process to a user pre-defined model. By offering this, users have greater freedom to define their own mappings and models, creating new semantic web information systems by integrating their datasets. These can be imported from a semantic web compliant format such as Resource Description Format or even from some miscellaneous source formats, such as tabular files or spreadsheets. In the spreadsheet case, the user has access to a set of tools to normalize the data and the possibility to create mappings by joining column data. These mappings are further converted into simple and connected triple patterns that are redirected to

the knowledge base through the API. This feature enables the transition from several spreadsheet formats to the semantic level by using a transparent layer and a user-friendly web-based interface (Figure 6.3).

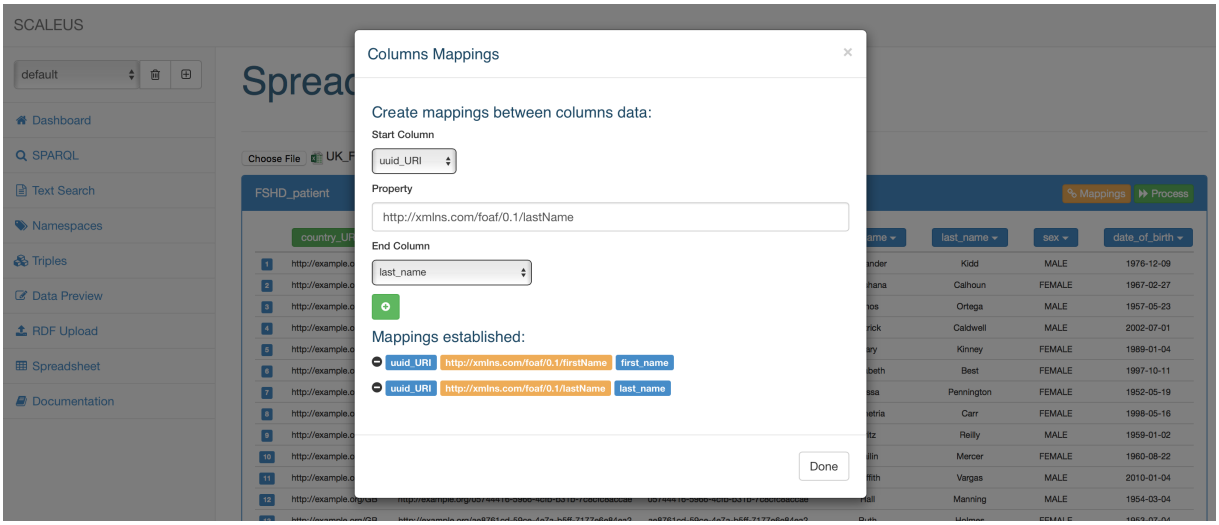


Figure 6.3: Web-based interface for multiple spreadsheet integration.

### 6.1.3 Inference support

As more complex and sophisticated ontologies are developed, the greater the need to develop systems to perform efficient queries that handle inference and deductive reasoning. Inference support indicates whether the proposed approach allows arbitrary deduction rules for inferring new knowledge, rather than the simple recording of facts. The inference on the SW is one of the most useful tools to enhance data integration quality, by discovering new relationships, automatically analyzing the content of the data, or managing knowledge on the Web in general. Hence, the inference ability is a key feature of many triplestores. However, while most triplestores' capabilities are limited to performing basic RDFS inference, only a few solutions, such as Sesame, allow specification of user-defined rules that can generate new relationships from existing triples, and therefore increase reasoning capabilities by inferring or discovering additional facts about the stored data.

In our system, we provide two different types of inference, one based on the RDFS classes and properties and another based on user-defined rules. The first supports the RDFS axioms and entailments described by the RDF Core working group, including the transitive closure of *subPropertyOf* and *subClassOf* relations and the *domain* or

*range* entailments. In this type of approach, we can infer facts without any pre-defined configuration, for instance, we can simply determine that: if the prototype for a class $A$ can be deduced as being a member of class $B$, then we conclude that $A$ is a *rdfs:subClassOf* $B$. The last implements a general rule-based inference mechanism giving more freedom to establish desired constraints. The rule-based inference mechanism can run in two modes: forward and backward chaining. The forward chaining engine is based on the standard RETE algorithm [189] and the backward chaining on logic programming, with similar execution to the Prolog engines. By using one of these modes, we can simply compose inference rules that, for instance, create new connections from our knowledge base such as: if $A$ is related with $B$ and $B$ is related with $C$ then we can assume that $A$ is related with $C$. This case can be translated to the system syntax by using the Dublin Core metadata [174] and the following simple rule:

```
[rule1:(?A dc:related ?B), (?B dc:related ?C) -> (?A dc:related ?C)]
```

Thus, an inference rule is defined by a list of body terms (premises), a list of head terms (conclusions), and an optional rule name and optional direction (chaining strategy). Each term is either a triple pattern, an extended triple pattern, or a call to a built-in primitive. The system also accepts a combination of a set of rules. This Scaleus rule-based inference feature works through real-time SPARQL queries over the selected dataset graph (Figure 6.4).

### 6.1.4 Text search index

This extension allows the combination of SPARQL query mechanisms and free-text search. In other words, it offers the ability to perform free-text searches within SPARQL queries. By using this extension, literals are tagged and indexed by an Apache Lucene (`http://lucene.apache.org`) engine. This combined solution allows faster search on object literals than just relying on use of the SPARQL match or filter expressions. Essentially, the text index is used to provide a reverse index mapping query strings to URIs. When data are loaded, any properties matching the description cause an entry to be added from analyzed text to the triple object and mapping to the subject. To retrieve the related subject, users can specify the target property and the respective string to search (native Lucene query language can be used). A separated Lucene index is created with each new dataset, storing all relevant information to perform free text search. This configuration is

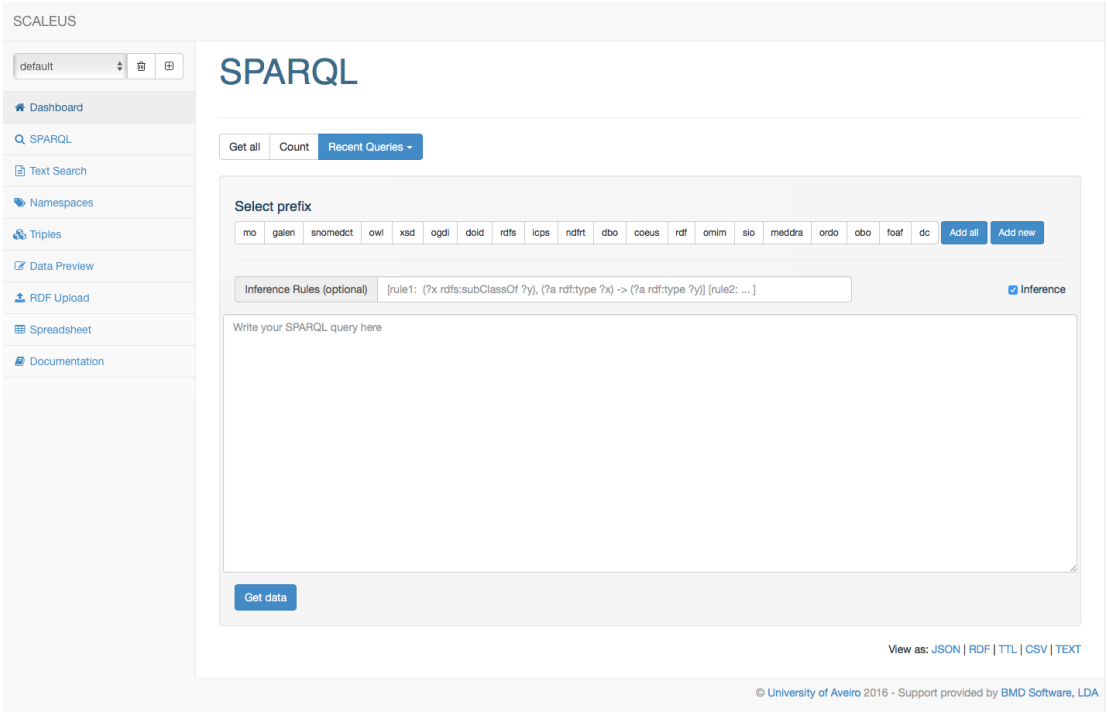


Figure 6.4: Scaleus SPARQL query engine supporting both RDFS and rule-based real-time inference.

passed to a Lucene IndexWriter structure for indexing. Furthermore, the text is examined by a text analyzer (i.e. the default Lucene *StandardAnalyzer*) that splits it into tokens and executes additional transformations such as the removal of stop words. The stemming and lemmatization processes are also applied to decrease inflectional forms and occasionally derivationally related formats of a word to a common base form. In this way, this extension takes advantage of all Lucene search library features. A web interface is provided to use this feature (Figure 6.2, Text Search option) and no initial configuration is needed.

## 6.2 Results

Scaleus aggregates several semantic-based mechanisms in a single package, providing the easy creation of new semantic web applications. The deployment of these services is suitable in a variety of life science scenarios. Nonetheless, the first use case implementation was focused on the rare disease domain.

### 6.2.1 Case Study

A common challenge in the field of rare diseases is the lack and complex nature of data. Studies are normally rare, heterogeneous, and distributed over different research centres and clinical labs. To allow identification of potential treatments, researchers need not only to identify a proper patient's cohort, but also to collect and combine relevant data from a wide range of repositories so that statistical significance can be obtained at the end. Nevertheless, as rare disease data is usually collected and maintained by different stakeholders in diverse and dedicated data warehouses, they are typically non-interoperable. To tackle these issues, we developed a Scaleus-based demonstrator that supports cross-resource queries over traditional rare disease resources, including biobanks (biological sample data), patient registries, genomic data and public repositories of biological relations. The demonstrator application (available at `http://bioinformatics.ua.pt/rd-connect-demo/`) was built in the context of the RD-Connect project [75] through an ELIXIR (`www.elixir-europe.org`) implementation study. In particular, the demonstrator enables queries across rare disease resources related to the "Ring14 syndrome", a very rare chromosomal abnormality [190], following the FAIR data principles [191]. It uses real resource metadata and data types, but the actual data have been obfuscated for the demonstrator proof-of-principle. According to the involved partners, Scaleus was essential in the provision of 3 main features: 1) the easy setup mechanism of a SW infrastructure without requiring demanding configurations; 2) easy data migration, using both simplified API and Graphical User Interface (GUI); and 3) the availablility of "out-of-the-box" key semantic services.

With the adoption of our approach during the semantic translation of registries, such as ID-CARD (`http://catalogue.rd-connect.eu`), general questions like "Which registries have data on patients with a diagnosis Ring14?", or more specific questions like "Give me blood specimens for patients that have a short neck" can now feasibly be answered. Furthermore, the demonstrator adopts a simple, tailored software platform, which offers a user-friendly query interface that sends automatically-generated SPARQL queries to the Scaleus middleware. First, users have to select a query template from a list, and subsequently, the interface offers a dropdown autocomplete widget to select specific values for required parameters. Then the system automatically generates a query that is solved by a Scaleus instance. Results are shown in a table that is marked with links to additional information. All of this is possible without requiring technical skills or even knowledge about SPARQL query language.

### 6.2.2 Evaluation

The platform assessment was performed through two distinct methodologies. Firstly, we measured the system performance against similar solutions. Secondly, we performed a user evaluation to validate implemented system features.

Scaleus was developed based on a native triplestore to offer fast query performance and improved scalability for the integrated datasets. In this case, we evaluate and compare four systems' query performance by analyzing two datasets of different sizes. The first dataset is the Orphanet [65] with around 400.000 triples. This dataset contains information on rare diseases and orphan drugs aiming to contribute to improvement of the diagnosis, care and treatment of patients with rare diseases. The second dataset is the GO Annotation (GOA) [92] with around 100.000.000 triples, containing high-quality Gene Ontology (GO) [192] annotations to proteins in the UniProt knowledge base [193] and International Protein Index (IPI) [194]. To evaluate the query performance, we used a load testing tool called locust (`http://locust.io`). With this open-source framework we simulate users' behaviour exploiting a suitable rate of 100 requests per second. Different queries were used to eliminate possible database cache mechanisms. Figure 6.5 shows the maximum response time for 50%, 75% and 95% of the requests. The results obtained show that our solution has greater performance in the majority of cases compared to the tested systems (i.e. Fuseki, Blazegraph and COEUS). The dataset dimension only delays our system response time by 2 milliseconds, if we observe 95% of the requests. In 75% of requests performed, our solution response time is not affected with either dataset. Being the only RDBMS-backed triplestore tested, COEUS achieved a maximum response of 44 milliseconds using the Orphanet dataset.

Regarding end-users' evaluation, we collected users' feedback during a hackathon organized for that propose. During one day, 10 participants (including biologists, bioinformatics researchers and software engineers) from 4 different European institutions were enrolled in the meeting. Thereafter, we conducted a survey to evaluate the user's satisfaction regarding several topics, such as the quality of documentation, setup effort, relevance of the implemented features (e.g. inference support, spreadsheet integration, text search, etc.), system's global performance, and the ability to access distributed data and overcome interoperability issues. These topics, were evaluated on a *Likert* scale, ranging from 1 (i.e. poor) to 5 (i.e. excellent). Figure 6.6 shows the results of this assessment, highlighting that the solution is very simple to instantiate and use for most users. According to this evaluation, features such as *Inference* support, *Spreadsheet* integration and *Text*

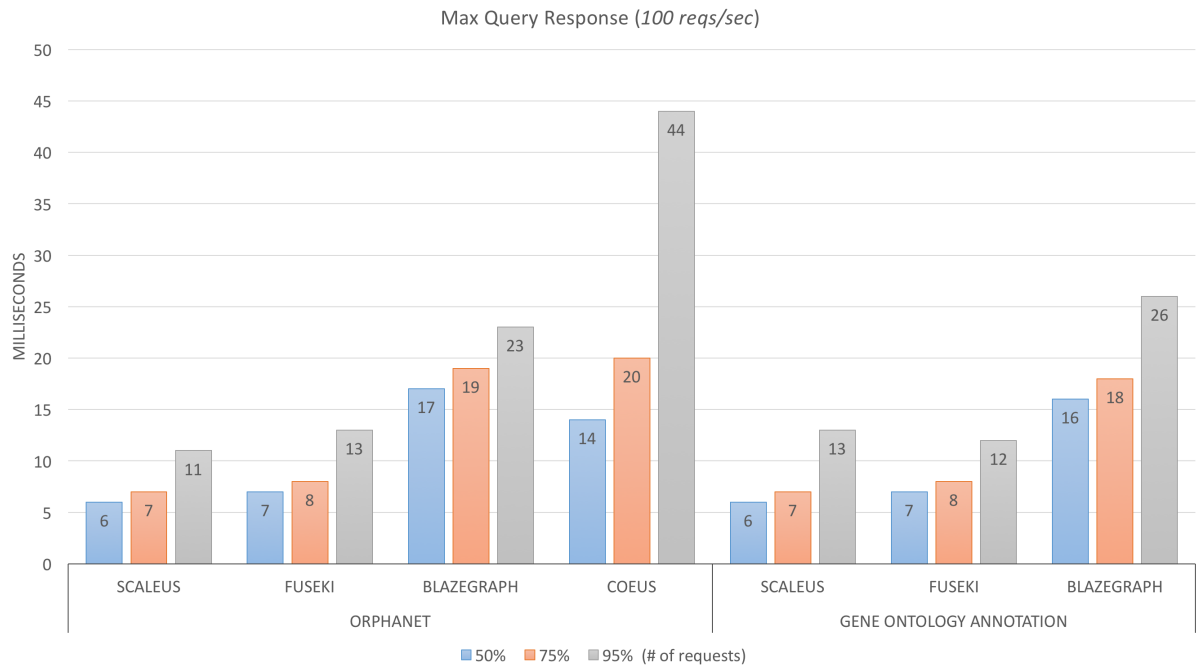


Figure 6.5: SPARQL query performance distribution evaluating four different triplestores' time response (lower is better) and showing the influence of two datasets of different sizes.

*Search* were demonstrated to be crucial for this type of system. Regarding the ability to access distributed data and solve interoperability issues, the users generally considered Scaleus as a very reliable platform to overcome these challenges. In terms of the overall rating, all users ranked Scaleus as *very good* or *excellent*, which suggests the tool is well accepted for the creation of new semantically interoperable repositories.

## 6.3 Discussion

Biomedical translational research requires technical infrastructures to deal with assorted data sources. By adopting SW technologies, we are limitless in exploring these data and shaping associations with external resources, avoiding traditional interoperability problems. Due to these characteristics, several data repositories and systems are gradually adopting semantic features, contributing to the comprehensive network being established across the research community. Despite this initial paradigm shift, few solutions provide easy, streamlined migration and deployment. This influences the delay in adapting existing datasets and applications to the semantic web environment.

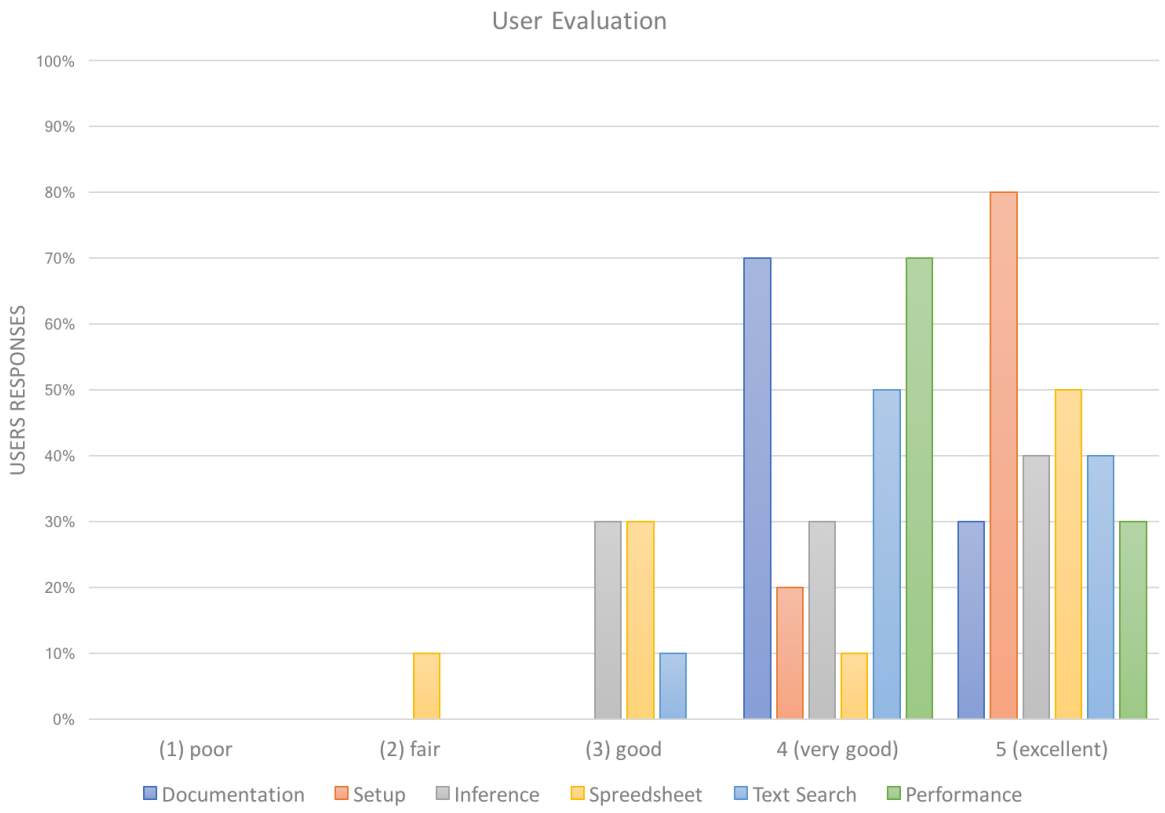


Figure 6.6: User evaluation overview.

In contrast, our solution was designed to offer an easy installation process with painless configuration, empowering life science applications with a streamlined semantic web deployment tool. This results in a flexible transition from traditional systems to an information system sustained by a fully semantic software stack. Furthermore, the majority of semantic conversion strategies depend on enabling mappings from relational connections, through real-time access such as D2R server [113], or based on pre-defined data integration ontologies such as COEUS [20]. These advances denote outstanding mapping algorithms, but are unsatisfactory when the goal is to provide a fast, customizable and native solution. Furthermore, the existence of fixed models for publishing datasets on the web is not suitable for most large-scale research projects as they need flexibility to test and spread their own prototype solutions. For those cases, Scaleus represents a better option as it can be used with any existing ontology and afterwards completed with information from traditional spreadsheets, for instance. If a custom integration is needed, the simplified REST API offers total freedom to insert new triples. This greatly increases adoption by biologists

or bioinformatics researchers since they do not need to acquire complex skills to deploy a semantic-based infrastructure.

Although Scaleus packs critical semantic features in a single system, its use cannot be separated, for instance, from data normalization tools. In other words, it can provide the foundation for the integration and distribution of conventional datasets but, in some specific cases, the alignment of annotation tools such as SORTA [195] or EGAS [119] with human experts' knowledge is required to map unstructured data beforehand. By offering robust data standardization methods, these semi-automated tools provide a perfect fit to be used in conjunction with our solution. In a sense, Scaleus provides essential baseline features to migrate to an empowered infrastructure exploiting a quicker path to publishing datasets through modern technologies.

## 6.4 Summary

In the past decades, biomedical research has generated a vast amount of miscellaneous information. Exploring these data is vital, and therefore, current computational systems need to be adjusted to assimilate and integrate this diversification.

The semantic web concept arises as a suitable environment for solving, at the same time, most data heterogeneity and interoperability challenges. Despite the evident features, transition to this paradigm and the respective establishment of semantic-based innovative bioinformatics tools is being delayed due to the lack of simplified and rapid migration solutions.

In this context, we developed Scaleus, a web-based open-source data migration tool to foster the adoption of semantic web technologies. Whilst it does not aim to provide a complete replacement infrastructure for existent systems, its side-by-side use offers a multitude of semantic features such as knowledge, inference rules and query federation of the available data. By delivering semantically-enhanced results, Scaleus dramatically increases the overall performance seamlessly across semantic networks, delivering a baseline foundation for the creation of sharable and interoperable bioinformatics platforms.

Summarizing, our solution enables the fast deployment of new semantic-based information systems by including, in a single package, the essential tools needed to contribute to the knowledge federation layer being established across life science research.

# Chapter 7

# Conclusions and future directions

Data integration is a key topic in the areas of computer science and biomedical informatics. Current computational solutions try to support data-intensive tasks to take advantage of the increasing information generated by both research and clinical practice. Nevertheless, data is growing at an unprecedented scale, in size and complexity, posing challenges for scientific innovation. Indeed, available infrastructures are not ready to assimilate and standardize the current nature of biomedical resources, delaying knowledge discovery advances. This creates a good opportunity to investigate novel methods towards an interoperable biomedical data network.

This document reports our efforts to tackle these challenges, providing enhanced methodologies to deal with the overwhelming data sources across the biomedical community.

## 7.1 Outcomes

We introduced our research discussing current data integration challenges at different levels and arguing about the benefits of making data interoperable across research and clinical institutions. Moreover, we examined available state-of-the-art strategies, identifying the gaps in existing methods and proposing enhanced solutions. Throughout this research process, several outcomes were achieved, contributing to the scientific endeavor in the distinct areas of knowledge.

The first contribution, the **Linked Registries** platform [18], was developed to simplify networking between data centres. The solution presented offers an opportunity to access patients' distributed data in a common web platform, supporting semantic data

representation, integrated access and querying. The connection between integrated patient registries using SW technologies allows federated inquires through multiple instances. Moreover, this contribution highlights our involvement in the RD-Connect project (`http://rd-connect.eu`), innovating in the creation of interoperable European-wide rare disease cohorts.

As a second contribution, we highlight the development of the second version of **COEUS** [19]. The level of automation introduced in this open-source platform makes the integration of data from multiple resources feasible, requiring minimal knowledge of the technologies involved. This significantly increases the usability and applicability of developed algorithms across the scientific community. In addition, this new version offers fast and efficient knowledge summarization, following the nanopublication standard. This enhances current scripting methodologies, augmenting the ability of non-informatician researchers to produce nanopublications.

The third contribution is focused on the production of information extraction workflows. Supported by the **Ann2RDF** tool [31], we introduced an interoperable architecture to unify text-mining outcomes and to enable proper exploration through multiple semantic-based interfaces [30]. This allows the harmonization of heterogeneous annotation results, enabling the easy creation of semantic-powered databases. The possibility to use and combine text-mined information from different and independent annotations is one of the main features of the implemented solution. The ability to process annotations from several, miscellaneous annotation formats benefits accessibility methods, allowing the integration of heterogeneous formats into a common and interoperable model.

Lastly, we underline the development of **SCALEUS**, a semantic web migration tool [32]. The lack of simplified and rapid migration solutions in the life sciences delays existing datasets and applications' adaptation to the semantic web environment. In this situation, our tool plays an important role, empowering life science applications with a streamlined semantic web deployment tool. The implemented solution facilitates the creation of new semantic web applications from scratch, presenting several semantic services for data integration and management. The package includes a native and high-performance triplestore supporting multiple independent datasets and a SPARQL query engine, supporting real-time inference mechanisms and optimized text searches over the knowledge base. The developed solution can be deployed in a large number of scenarios, offering wide applicability for the interdisciplinary field of biomedicine. This is a major outcome of our research, as it is possible to reuse our solutions to advance new specialized

applications.

## 7.2 Future work

This thesis introduced novel methods and architectures that can be deeply explored in different scenarios. Additionally, we can identify some research lines for further exploration in the future.

A first challenge that can be addressed is related to the enhancement of web interfaces for data exploration. During our research, we developed several services for knowledge bases' integration and access. Some of these services, such as SPARQL, require knowledge about their main functions to be effectively used. This is a critical point for the adoption of semantic web information systems, and further research should be more concerned with user usability and functionality aspects.

Another future line of research could be the exploration of data analytics tools. Semantic web-based repositories make data interoperable, offering additional and important insights into the data collected. However, most of the available systems are oriented to information retrieval, lacking user-assisted methods for data analysis. Adding more complex data analysis features will offer additional opportunities for exploring these valuable data networks.

# References


[1] J. Watson, "The human genome project: past, present, and future", *Science*, vol. 248, no. 4951, pp. 44–49, Apr. 1990.

[2] L. Hunter and K. B. Cohen, "Biomedical language processing: what's beyond pubmed?", *Molecular cell*, vol. 21, no. 5, pp. 589–94, Mar. 2006.

[3] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: current progress.", *Briefings in bioinformatics*, vol. 8, no. 5, pp. 358–75, Sep. 2007.

[4] D. Campos, S. Matos, and J. Oliveira, "Current methodologies for biomedical named entity recognition", *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*, pp. 839–868, 2012.

[5] B. Alex, C. Grover, and B. Haddow, "Assisted curation: does text mining really help?.", *Pacific Symposium on Biocomputing*, vol. 13, 2008.

[6] H. Dai, Y. Chang, R. Tsai, and W. Hsu, "New challenges for biological text-mining in the next decade", *Journal of Computer Science and Technology*, vol. 25.1, pp. 169–179, 2010.

[7] K. Döring, B. A. Grüning, K. K. Telukunta, P. Thomas, and S. Günther, "Pubmedportable: a framework for supporting the development of text mining applications", *PLOS ONE*, vol. 11, no. 10, R. Guralnick, Ed., e0163794, Oct. 2016.

[8] D. Campos, S. Matos, and J. Oliveira, "Neji: a tool for heterogeneous biomedical concept identification", *Proceedings of BioLINK SIG*, 2013.

[9] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology", *Trends in Genetics*, vol. 30, no. 9, pp. 418–426, Aug. 2014.

[10] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, *et al.*, "Data integration in the era of omics: current and future challenges.", En, *BMC systems biology*, vol. 8 Suppl 2, no. 2, p. I1, Jan. 2014.

[11] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web", *Scientific american*, vol. 284.5, pp. 28–37, 2001.

[12] T. Passin, *Explorer's Guide to the Semantic Web*. 2004, ISBN: 1932394206.

[13] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2rdf: towards a mashup to build bioinformatics knowledge systems.", *Journal of biomedical informatics*, vol. 41, no. 5, pp. 706–16, Oct. 2008.

[14] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, *et al.*, "The ebi rdf platform: linked open data for the life sciences", *Bioinformatics*, vol. 30, no. 9, pp. 1338–1339, 2014.

[15] P. Sernadela, P. Lopes, and J. L. Oliveira, "A knowledge federation architecture for rare disease patient registries and biobanks", *Journal of Information Systems Engineering & Management*, vol. 1, no. 1, pp. 83–90, 2016.

[16] A. Freitas, E. Curry, J. G. Oliveira, and S. O'Riain, "Querying heterogeneous datasets on the linked data web: challenges, approaches, and trends", *IEEE Internet Computing*, vol. 16, no. 1, pp. 24–33, Jan. 2012.

[17] P. Lopes, P. Sernadela, and J. L. Oliveira, "Towards a knowledge federation of linked patient registries", in *2015 10th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, Jun. 2015, pp. 1–5, ISBN: 978-9-8998-4345-5.

[18] P. Sernadela, L. González-Castro, C. Carta, E. van der Horst, P. Lopes, *et al.*, "Linked registries: Connecting rare diseases patient registries through a semantic web layer", *BioMed Research International*, vol. 2017, pp. 1–13, Oct. 2017.

[19] P. Sernadela and J. L. Oliveira, "Coeus 2.0: an automated platform to integrate and publish biomedical data as nanopublications", *IET Software*, vol. 11, Dec. 2017.

[20] P. Lopes and J. L. Oliveira, "Coeus: "semantic web in a box" for biomedical applications.", *Journal of biomedical semantics*, vol. 3, no. 1, p. 11, Jan. 2012.

[21] P. Lopes, P. Sernadela, and J. L. Oliveira, "Exploring nanopublishing with coeus", in *6th International Workshop on Semantic Web Applications and Tools for Life Sciences*, vol. 1114, Edinburgh: CEUR-WS, 2013.

[22] A. Roos, S. Beltran, D. Piscia, S. Laurie, J. Protasio, *et al.*, "Rd-connect: data sharing and analysis for rare disease research within the integrated platform and through ga4gh beacon and matchmaker exchange", *Neuromuscular Disorders*, vol. 26, S160–S161, Oct. 2016.

[23] R. Kaliyaperumal, P. A. C. T'Hoen, Z. Tatum, M. Thompson, E. Van Der Horst, *et al.*, "Genome annotation using nanopublications: an approach to interoperability of genetic data", in *CEUR Workshop Proceedings*, vol. 1320, CEUR-WS, 2014.

[24] P. Sernadela, A. Pereira, and R. Rossetti, "Disim: ontology-driven simulation of biomedical data integration tasks", in *2015 10th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, Jun. 2015, pp. 1–4, ISBN: 978-9-8998-4345-5.

[25] P. Sernadela and J. L. Oliveira, "Automated nanopublications generation from biomedical literature", in *2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG)*, IEEE, 2017, pp. 1–4, ISBN: 978-1-5090-4801-4.

[26] P. Sernadela, E. van der Horst, M. Thompson, P. Lopes, M. Roos, *et al.*, "A nanopublishing architecture for biomedical data", in *8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*, vol. 294, Springer International Publishing, 2014, pp. 277–284.

[27] P. Sernadela, P. Lopes, and J. L. Oliveira, "Exploring nanopublications integration in pharmacovigilance scenarios", in *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, Lisbon, Portugal: IEEE, Oct. 2013, pp. 728–730, ISBN: 978-1-4673-5801-9.

[28] P. Sernadela, P. Lopes, D. Campos, S. Matos, and J. L. Oliveira, "A semantic layer for unifying and exploring biomedical document curation results", in *Bioinformatics and Biomedical Engineering. Springer International Publishing.* 2015, pp. 8–17.

[29] E. Monteiro, P. Sernadela, S. Matos, C. Costa, and J. L. Oliveira, "Semantic knowledge base construction from radiology reports", in *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, SCITEPRESS - Science, 2016, pp. 345–352, ISBN: 978-989-758-170-0.

[30] P. Sernadela and J. L. Oliveira, "A semantic-based workflow for biomedical literature annotation", *Database*, vol. 2017, bax088, Jan. 2017.

[31] P. Sernadela, S. Matos, and J. L. Oliveira, "Ann2rdf: moving annotations to semantic web", in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications &Services - iiWAS '15*, New York, New York, USA: ACM Press, Dec. 2015, pp. 1–5, ISBN: 9781450334914.

[32] P. Sernadela, L. González-Castro, and J. L. Oliveira, "Scaleus: semantic web services integration for biomedical applications", *Journal of Medical Systems*, vol. 41, no. 4, p. 54, Apr. 2017.

[33] D. J. Rigden, X. M. Fernández-Suárez, and M. Y. Galperin, "The 2016 database issue of nucleic acids research and an updated molecular biology database collection", *Nucleic Acids Research*, vol. 44, no. D1, pp. D1–D6, Jan. 2016.

[34] D. Zou, L. Ma, J. Yu, and Z. Zhang, "Biological databases for human research", *Genomics, Proteomics & Bioinformatics*, vol. 13, no. 1, pp. 55–63, Feb. 2015.

[35] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, *et al.*, "The comparative toxicogenomics database: update 2017", *Nucleic Acids Research*, vol. 45, no. D1, p. D972, 2017.

[36] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, *et al.*, "Drugbank 4.0: shedding new light on drug metabolism", *Nucleic Acids Research*, vol. 42, no. D1, p. D1091, 2014.

[37] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi", *Nucleic Acids Research*, vol. 33, no. Suppl 1, p. D54, 2005.

[38] P. Artimo, M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi, *et al.*, "Expasy: sib bioinformatics resource portal", *Nucleic Acids Research*, vol. 40, no. W1, W597, 2012.

[39] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, *et al.*, "Genbank", *Nucleic Acids Research*, vol. 41, no. D1, p. D36, 2013.

[40] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, "Genenames.org: the hgnc resources in 2015", *Nucleic Acids Research*, vol. 43, no. D1, p. D1079, 2015.

[41] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, *et al.*, "Hmdb 3.0–the human metabolome database in 2013", *Nucleic Acids Research*, vol. 41, no. D1, p. D801, 2013.

[42] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: new perspectives on genomes, pathways, diseases and drugs", *Nucleic Acids Research*, vol. 45, no. D1, p. D353, 2017.

[43] R. Fescharek, J. Kübler, U. Elsasser, M. Frank, and P. Güthlein, "Medical dictionary for regulatory activities (meddra)", *International Journal of Pharmaceutical Medicine*, vol. 18, no. 5, pp. 259–269, 2014.

[44] C. E. Lipscomb, "Medical subject headings (mesh)", eng, *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–266, 2000.

[45] L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, *et al.*, "The ncbi biosystems database", *Nucleic Acids Research*, vol. 38, no. Suppl 1, p. D492, 2010.

[46] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders", *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D514–D517, 2005.

[47] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, *et al.*, "Pharmacogenomics knowledge for personalized medicine", *Clinical Pharmacology & Therapeutics*, vol. 92, no. 4, pp. 414–417, 2012.

[48] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, *et al.*, "The protein data bank", *Nucleic Acids Research*, vol. 28, no. 1, p. 235, 2000.

[49] S. J. Nelson, K. Zeng, J. Kilbourne, T. Powell, and R. Moore, "Normalized names for clinical drugs: rxnorm at 6 years", *Journal of the American Medical Informatics Association*, vol. 18, no. 4, p. 441, 2011.

[50] R. Cornet and N. de Keizer, "Forty years of snomed: a literature review", *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, S2, 2008.

[51] P. Wexler, "Toxnet: an evolving web resource for toxicology and environmental health information", *Toxicology*, vol. 157, no. 1–2, pp. 3–10, 2001.

[52] The UniProt Consortium, "Uniprot: the universal protein knowledgebase", *Nucleic Acids Research*, vol. 45, no. D1, p. D158, 2017.

[53] EU, *European commission's communication on rare diseases: europe's challenge*, 2008. [Online]. Available: `http://ec.europa.eu/health/ph_threats/non_com/docs/rare_com_en.pdf` (visited on 05/30/2017).

[54] E. Seoane-Vazquez, R. Rodriguez-Monguio, S. L. Szeinbach, and J. Visaria, "Incentives for orphan drug research and development in the united states", *Orphanet journal of rare diseases*, vol. 3, p. 33, 2008.

[55] A. Schieppati, J. I. Henter, E. Daina, and A. Aperia, "Why rare diseases are an important medical and social issue", *The Lancet*, vol. 371, no. 9629, pp. 2039–2041, 2008.

[56] D. N. Cooper, J. Chen, E. V. Ball, K. Howells, M. Mort, *et al.*, "Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics", *Human mutation*, vol. 31, no. 6, pp. 631–655, 2010.

[57] S. Aymé and J. Schmidtke, "Networking for rare diseases: a necessity for europe", *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, vol. 50, no. 12, pp. 1477–1483, 2007.

[58] Orphanet, *Orphanet report series: rare disease registries in europe*, 2016. [Online]. Available: `http://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf` (visited on 05/30/2017).

[59] K. M. Boycott, M. R. Vanstone, D. E. Bulman, and A. E. MacKenzie, "Rare-disease genetics in the era of next-generation sequencing: discovery to translation", *Nat Rev Genet*, vol. 14, no. 10, pp. 681–691, 2013.

[60] J. K. Aronson, "Rare diseases and orphan drugs", *British Journal of Clinical Pharmacology*, vol. 61, no. 3, pp. 243–245, 2006.

[61] M. Wastfelt, B. Fadeel, and J. I. Henter, "A journey of hope: lessons learned from studies on rare diseases and orphan drugs", *Journal of Internal Medicine*, vol. 260, no. 1, pp. 1–10, 2006.

[62] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, *et al.*, "Human gene mutation database (hgmd®): 2003 update", *Human mutation*, vol. 21, no. 6, pp. 577–581, 2003.

[63] M. Via, C. Gignoux, and E. G. Burchard, "The 1000 genomes project: new opportunities for research and social challenges", *Genome Med*, vol. 2, no. 3, 2010.

[64] B. Mons, H. van Haagen, C. Chichester, P.-B. '. Hoen, J. T. den Dunnen, *et al.*, "The value of data.", en, *Nature genetics*, vol. 43, no. 4, pp. 281–3, Apr. 2011.

[65] A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, *et al.*, "Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users", *Human Mutation*, vol. 33, no. 5, pp. 803–808, 2012.

[66] P. Lopes and J. L. Oliveira, "An innovative portal for rare genetic diseases research: the semantic diseasecard.", *Journal of biomedical informatics*, Aug. 2013.

[67] C. L. Bladen, R. Thompson, J. M. Jackson, C. Garland, C. Wegel, *et al.*, "Mapping the differences in care for 5,000 spinal muscular atrophy patients, a survey of 24 national registries in north america, australasia and europe", *Journal of neurology*, vol. 261, no. 1, pp. 152–163, 2014.

[68] R. Somerville, A. Jackson, S. Zhou, G. Fletcher, and P. Fitzpatrick, "Non-pulmonary chronic diseases in adults with cystic fibrosis: analysis of data from the cystic fibrosis registry", *Irish medical journal*, 2013.

[69] B. Martin, M. S. Schechter, A. Jaffe, P. Cooper, S. C. Bell, *et al.*, "Comparison of the us and australian cystic fibrosis registries: the impact of newborn screening", *Pediatrics*, vol. 129, no. 2, e348–e355, 2012.

[70] A. Sárközy, K. Bushby, C. Béroud, and H. Lochmüller, "157th enmc international workshop: patient registries for rare, inherited muscular disorders 25–27 january 2008 naarden, the netherlands", *Neuromuscular Disorders*, vol. 18, no. 12, pp. 997–1001, 2008.

[71] C. L. Bladen, K. Rafferty, V. Straub, S. Monges, A. Moresco, *et al.*, "The treat-nmd duchenne muscular dystrophy registries: conception, design, and utilization by industry and academia", *Human mutation*, vol. 34, no. 11, pp. 1449–1457, 2013.

[72] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology", *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[73] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang, "Snomed clinical terms: overview of the development process and project status.", *Proceedings of the AMIA Symposium*, pp. 662–6, Jan. 2001.

[74] P. N. Robinson and S. Mundlos, "The human phenotype ontology", *Clinical genetics*, vol. 77, no. 6, pp. 525–534, 2010.

[75] R. Thompson, L. Johnston, D. Taruscio, L. Monaco, C. Béroud, *et al.*, "Rd-connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research.", *Journal of general internal medicine*, Jul. 2014.

[76] A. A. Philippakis, D. R. Azzariti, S. Beltran, A. J. Brookes, C. A. Brownstein, *et al.*, "The matchmaker exchange: a platform for rare disease gene discovery", *Human Mutation*, vol. 36, no. 10, pp. 915–921, Oct. 2015.

[77] J. Hebeler, M. Fisher, R. Blace, A. Perez-Lopez, and M. Dean, *Semantic Web Programming*. Wiley, 2009, p. 652, ISBN: 978-0-470-41801-7.

[78] P. Banerjee, R. Friedrich, C. Bash, P. Goldsack, B. Huberman, *et al.*, "Everything as a service: powering the new information economy", *Computer*, vol. 44, no. 3, pp. 36–43, Mar. 2011.

[79] T. Berners-Lee, *Design issues: linked data*, 2006. [Online]. Available: `https://www.w3.org/DesignIssues/LinkedData.html` (visited on 07/28/2014).

[80] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The meaningful use of big data", *ACM SIGMOD Record*, vol. 40, no. 4, p. 56, Jan. 2012.

[81] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far", *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[82] J. Lehmann, R. Isele, and M. Jakob, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia", *Semantic Web*, 2014.

[83] L. Masinter, T. Berners-Lee, and R. Fielding, "Uniform resource identifier (uri): generic syntax", 2005.

[84] E. Miller, "An introduction to the resource description framework", *Journal of Library Administration*, 2001.

[85] G. Klyne and J. Carroll, "Resource description framework (rdf): concepts and abstract syntax. w3c recommendation, 2004", *World Wide Web Consortium*, 2004.

[86] M. Uschold and M. Gruninger, "Ontologies: principles, methods and applications", *Knowledge engineering review*, 1996.

[87] P. Patel-Schneider, P. Hayes, and I. Horrocks, "Owl web ontology language semantics and abstract syntax", *W3C recommendation*, 2004.

[88] J. Bard, S. Y. Rhee, and M. Ashburner, "An ontology for cell types", *Genome Biology*, vol. 6, no. 2, R21, 2005.

[89] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, *et al.*, "The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013", *Nucleic Acids Research*, vol. 41, no. D1, pp. D456–D463, Nov. 2012.

[90] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, *et al.*, "Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data", *Nucleic Acids Research*, vol. 43, no. D1, pp. D1071–D1078, 2015.

[91] M. Herrero-Zazo, I. Segura-Bedmar, J. Hastings, and P. Martínez, "Dinto: using owl ontologies and swrl rules to infer drug–drug interactions and their mechanisms", *Journal of Chemical Information and Modeling*, vol. 55, no. 8, pp. 1698–1707, Aug. 2015.

[92] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, *et al.*, "The goa database in 2009–an integrated gene ontology annotation resource", *Nucleic Acids Research*, vol. 37, no. Database, pp. D396–D403, Jan. 2009.

[93] D. A. Natale, C. N. Arighi, W. C. Barker, J. A. Blake, C. J. Bult, *et al.*, "The protein ontology: a structured representation of protein forms and complexes", *Nucleic Acids Research*, vol. 39, no. Database, pp. D539–D545, Jan. 2011.

[94] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, *et al.*, "The sequence ontology: a tool for the unification of genome annotations", *Genome Biology*, vol. 6, no. 5, R44, 2005.

[95] J. Pathak, R. Kiefer, and C. Chute, "Using semantic web technologies for cohort identification from electronic health records for clinical research", *AMIA Summits on Translational Science Proceedings*, 2012.

[96] C. David, C. Olivier, and B. Guillaume, "A survey of rdf storage approaches", *ARIMA Journal*, 2012.

[97] O. Erling, "Virtuoso, a hybrid rdbms/graph column store.", Tech. Rep., 2012.

[98] P. Lopes and J. L. Oliveira, "Coeus: a semantic web application framework", in *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*, ACM, 2011, pp. 66–73, ISBN: 1450310761.

[99] J. Broekstra, A. Kampman, and F. V. Harmelen, "Sesame: a generic architecture for storing and querying rdf and rdf schema", *The Semantic Web — ISWC 2002*, vol. 2342, pp. 54–68, 2002.

[100] J. Aasman, "Allegro graph: rdf triple database", *Cidade: Oakland Franz Incorporated*, 2006.

[101] G. E. Modoni, M. Sacco, and W. Terkaj, "A survey of rdf store solutions", in *2014 International Conference on Engineering, Technology and Innovation (ICE)*, IEEE, Jun. 2014, pp. 1–7, ISBN: 978-1-4799-4735-5.

[102] E. Prud'Hommeaux, A. Seaborne, E. Prud'Hommeaux, and A. Seaborne, "Sparql query language for rdf", *W3C recommendation*, vol. 15, 2008.

[103] N. A. Rakhmawati, J. Umbrich, M. Karnstedt, A. Hasnain, and M. Hausenblas, "Querying over federated sparql endpoints —a state of the art survey", Jun. 2013. arXiv: 1306.1723.

[104] C. Goble and R. Stevens, "State of the nation in data integration for bioinformatics.", *Journal of biomedical informatics*, vol. 41, no. 5, pp. 687–93, Oct. 2008.

[105] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies.", *Bioinformatics (Oxford, England)*, vol. 26, no. 4, pp. 445–55, Feb. 2010.

[106] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, "Bioinformatics challenges for personalized medicine.", *Bioinformatics (Oxford, England)*, vol. 27, no. 13, pp. 1741–8, Jul. 2011.

[107] N. Cannata, M. Schröder, R. Marangoni, and P. Romano, "A semantic web for bioinformatics: goals, tools, systems, applications.", *BMC bioinformatics*, vol. 9 Suppl 4, no. Suppl 4, S1, Jan. 2008.

[108] M. D. Wilkinson, B. P. Vandervalk, E. L. McCarthy, and L. McCarthy, "The semantic automated discovery and integration (sadi) web service design-pattern, api and reference implementation", *Journal of biomedical semantics*, vol. 2, no. 1, p. 8, Jan. 2011.

[109] H. Mohamed, Y. Jincai, and J. Qian, "Towards integration rules of mapping from relational databases to semantic web ontology", *Web Information Systems and Mining (WISM)*, 2010.

[110] W3C, "A survey of current approaches for mapping of relational databases to rdf', *W3C RDB2RDF Incubator Group Report*, 2009.

[111] G. Būmans and K. Čerāns, "Rdb2owl", in *Proceedings of the 6th International Conference on Semantic Systems - I-SEMANTICS '10*, New York, New York, USA: ACM Press, Sep. 2010, p. 1, ISBN: 9781450300148.

[112] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller, "Triplify", in *Proceedings of the 18th international conference on World wide web - WWW '09*, New York, New York, USA: ACM Press, Apr. 2009, p. 621, ISBN: 9781605584874.

[113] C. Bizer and R. Cyganiak, "D2r server-publishing relational databases on the semantic web", *5th International Semantic Web Conference*, 2006.

[114] D. Rebholz-Schuhmann, A. Oellrich, and R. Hoehndorf, "Text-mining solutions for biomedical research: enabling integrative biology", *Nature Reviews Genetics*, vol. 13, no. 12, pp. 829–839, Nov. 2012.

[115] R. Khare, R. Leaman, and Z. Lu, "Accessing biomedical literature in the current information landscape", in *PLoS ONE*, 2014, pp. 11–31.

[116] P. Stenetorp, S. Pyysalo, and G. Topić, "Brat: a web-based tool for nlp-assisted text annotation", *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107, 2012.

[117] D. Salgado, M. Krallinger, M. Depaule, E. Drula, A. V. Tendulkar, *et al.*, "Myminer: a web application for computer-assisted biocuration and text annotation.", *Bioinformatics (Oxford, England)*, vol. 28, no. 17, pp. 2285–7, Sep. 2012.

[118] R. Rak, A. Rowley, W. Black, and S. Ananiadou, "Argo: an integrative, interactive, text mining-based workbench supporting curation.", *Database : The journal of biological databases and curation*, vol. 2012, bas010, Jan. 2012.

[119] D. Campos, J. Lourenco, S. Matos, and J. L. Oliveira, "Egas: a collaborative and interactive document curation platform", *Database*, vol. 2014, Jun. 2014.

[120] K. Verspoor, K. B. Cohen, A. Lanfranchi, C. Warner, H. L. Johnson, *et al.*, "A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools.", *BMC bioinformatics*, vol. 13, no. 1, p. 207, Jan. 2012.

[121] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus–a semantically annotated corpus for bio-textmining", *Bioinformatics*, vol. 19, no. Suppl 1, pp. i180–i182, Jul. 2003.

[122] P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction.", *BMC bioinformatics*, vol. 10, no. 1, p. 349, Jan. 2009.

[123] A. J. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson, "Exploiting mesh indexing in medline to generate a data set for word sense disambiguation", *BMC Bioinformatics*, vol. 12, no. 1, p. 223, 2011.

[124] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction.", *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, no. 5, pp. 544–51, Jan. 2011.

[125] D. Campos and Q. Bui, "Trigner: automatically optimized biomedical event trigger recognition on scientific documents.", *Source code for biology and medicine*, 2014.

[126] K. Tomanek, J. Wermter, and U. Hahn, "A reappraisal of sentence and token splitting for life sciences documents", in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, IOS Press, 2007, pp. 524–528.

[127] N. Barrett and J. Weber-Jahnke, "Building a biomedical tokenizer using the token lattice design pattern and the adapted viterbi algorithm", *BMC Bioinformatics*, vol. 12, no. Suppl 3, S1, 2011.

[128] Y. He and M. Kayaalp, "A comparison of 13 tokenizers on medline", *MD: The Lister Hill National Center for Biomedical Communications*, 2006.

[129] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, *et al.*, "Developing a robust part-of-speech tagger for biomedical text", in *Proceedings of the 10th Panhellenic Conference on Advances in Informatics*, Springer Berlin Heidelberg, 2005, pp. 382–392.

[130] H. Liu, T. Christiansen, W. A. Baumgartner, and K. Verspoor, "Biolemmatizer: a lemmatization tool for morphological processing of biomedical text", *Journal of Biomedical Semantics*, vol. 3, no. 1, p. 3, 2012.

[131] N. Kang, E. M. van Mulligen, and J. A. Kors, "Comparing and combining chunkers of biomedical text.", *Journal of biomedical informatics*, vol. 44, no. 2, pp. 354–60, Apr. 2011.

[132] U. Leser and J. Hakenberg, "What makes a gene name? named entity recognition in the biomedical literature", *Briefings in Bioinformatics*, vol. 6, no. 4, pp. 357–369, Jan. 2005.

[133] R. Gaizauskas and G. Demetriou, "Protein structures and information extraction from biological texts: the pasta system", *Bioinformatics*, 2003.

[134] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: a machine learning approach.", *Bioinformatics (Oxford, England)*, vol. 20, no. 7, pp. 1178–90, May 2004.

[135] A. Jimeno-Yepes and A. Aronson, "Self-training and co-training in biomedical word sense disambiguation", *Proceedings of BioNLP 2011 Workshop*, 2011.

[136] A. Jimeno-Yepes and A. Aronson, "Knowledge-based biomedical word sense disambiguation: comparison of approaches", *BMC bioinformatics*, 2010.

[137] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. B. Kell, "Event extraction for systems biology by text mining the literature.", *Trends in biotechnology*, vol. 28, no. 7, pp. 381–90, Jul. 2010.

[138] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, *et al.*, "Biomedical text mining and its applications in cancer research.", *Journal of biomedical informatics*, vol. 46, no. 2, pp. 200–11, Apr. 2013.

[139] Q.-C. Bui, S. Katrenko, and P. M. A. Sloot, "A hybrid approach to extract protein-protein interactions.", *Bioinformatics (Oxford, England)*, vol. 27, no. 2, pp. 259–65, Jan. 2011.

[140] L. Tari, S. Anwar, S. Liang, J. Cai, and C. Baral, "Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism.", *Bioinformatics (Oxford, England)*, vol. 26, no. 18, pp. i547–53, Sep. 2010.

[141] T. C. Wiegers, A. Davis, K. B. Cohen, L. Hirschman, and C. J. Mattingly, "Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd)", *BMC Bioinformatics*, vol. 10, no. 1, p. 326, Oct. 2009.

[142] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of bionlp'09 shared task on event extraction", pp. 1–9, Jun. 2009.

[143] Q. Bui, E. V. Mulligen, D. Campos, and J. Kors, "A fast rule-based approach for biomedical event extraction", *ACL 2013*, 2013.

[144] A. Ben Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach", *Journal of Biomedical Semantics*, vol. 2, no. Suppl 5, S4, 2011.

[145] D. Rebholz-Schuhmann and H. Kirsch, "Iexml: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules", *SIG BioLink, ISMB*, 2006.

[146] D. C. Comeau, R. Islamaj Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, *et al.*, "Bioc: a minimalist approach to interoperability for biomedical text processing.", *Database : The journal of biological databases and curation*, vol. 2013, bat064, Jan. 2013.

[147] C. M. Machado, D. Rebholz-Schuhmann, A. T. Freitas, and F. M. Couto, "The semantic web in translational medicine: current applications and future directions", *Briefings in Bioinformatics*, vol. 16, no. 1, pp. 89–103, Jan. 2015.

[148] J. Laurila, N. Naderi, and R. Witte, "Algorithms and semantic infrastructure for mutation impact extraction and grounding", *BMC genomics*, vol. S24, 2010.

[149] A. Coulet, Y. Garten, M. Dumontier, R. B. Altman, M. A. Musen, *et al.*, "Integration and publication of heterogeneous text-mined relationships on the semantic web.", *Journal of biomedical semantics*, vol. 2 Suppl 2, S10, Jan. 2011.

[150] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight", in *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics '11*, New York, New York, USA: ACM Press, Sep. 2011, pp. 1–8, ISBN: 9781450306218.

[151] J. Kim and Y. Wang, "Pubannotation: a persistent and sharable corpus and annotation repository", *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistic*, 2012.

[152] R. Rak, R. T. Batista-Navarro, J. Carter, A. Rowley, and S. Ananiadou, "Processing biological literature with customizable web services supporting interoperable formats", *Database*, vol. 2014, bau064–bau064, Jul. 2014.

[153] N. R. Anderson, E. S. Lee, J. S. Brockenbrough, M. E. Minie, S. Fuller, *et al.*, "Issues in biomedical research data management and analysis: needs and barriers.", *Journal of the American Medical Informatics Association : JAMIA*, vol. 14, no. 4, pp. 478–88, Jan. 2007.

[154] G. P. Patrinos, D. N. Cooper, E. van Mulligen, V. Gkantouna, G. Tzimas, *et al.*, "Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain.", *Human mutation*, vol. 33, no. 11, pp. 1503–12, Nov. 2012.

[155] J. Velterop, "Nanopublications*: the future of coping with information overload.", *LOGOS: The Journal of the World Book Community*, 2010.

[156] B. Giardine, J. Borg, D. R. Higgs, K. R. Peterson, S. Philipsen, *et al.*, "Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach.", *Nature genetics*, vol. 43, no. 4, pp. 295–301, Apr. 2011.

[157] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nanopublication", *Information Services and Use*, 2010.

[158] L. Harland, "Open phacts: a semantic knowledge infrastructure for public and commercial drug discovery research", *Knowledge Engineering and Knowledge Management*, 2012.

[159] E. Mina and M. Thompson, "Nanopublications for exposing experimental data in the life-sciences: a huntington's disease case study", *Proc. 6th Int. Semant. Web Appl. Tools Life Sci. Work.(SWAT4LS 2013)*, 2013.

[160] C. Chichester, O. Karch, P. Gaudet, L. Lane, B. Mons, *et al.*, "Converting nextprot into linked data and nanopublications", *Semantic Web*, vol. 6, no. 2, pp. 147–153, 2015.

[161] E. Mina, M. Thompson, R. Kaliyaperumal, J. Zhao, E. van der Horst, *et al.*, "Nanopublications for exposing experimental data in the life-sciences: a huntington's disease case study", *Journal of Biomedical Semantics*, vol. 6, no. 1, p. 5, 2015.

[162] I. F. A. C. Fokkema, P. E. M. Taschner, G. C. P. Schaafsma, J. Celli, J. F. J. Laros, *et al.*, "Lovd v.2.0: the next generation in gene variant databases.", *Human mutation*, vol. 32, no. 5, pp. 557–63, May 2011.

[163] J. McCusker and T. Lebo, "Next generation cancer data discovery, access, and integration using prizms and nanopublications", *Data Integration in the Life Sciences*, pp. 105–112, 2013.

[164] T. Kuhn and M. Krauthammer, "Underspecified scientific claims in nanopublications", *ArXiv preprint arXiv:1209.1483*, 2012.

[165] T. Kuhn and P. Barbano, "Broadening the scope of nanopublications", *The Semantic Web: Semantics and Big Data*, pp. 487–501, 2013.

[166] B. Mons and J. Velterop, "Nano-publication in the e-science era", *Workshop on Semantic Web Applications in Scientific Discourse*, 2009.

[167] K. Belhajjame, O. Corcho, and D. Garijo, "Workflow-centric research objects: first class citizens in scholarly discourse", *Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web*, 2012.

[168] K. Belhajjame, J. Zhao, D. Garijo, K. Hettne, R. Palma, *et al.*, "The research object suite of ontologies: sharing and exchanging research data and methods on the open web", *ArXiv preprint arXiv: 1401.4307*, p. 20, Jan. 2014. arXiv: 1401.4307.

[169] M. Marshall, L. Post, M. Roos, and T. Breit, "Using semantic web tools to integrate experimental measurement data on our own terms", *On the Move to Meaningful Internet Systems*, 2006.

[170] S. P. Gardner, "Ontologies and semantic data integration", *Drug discovery today*, vol. 10, no. 14, pp. 1001–1007, 2005.

[171] C. Pasquier, "Biological data integration using semantic web technologies", *Biochimie*, vol. 90, no. 4, pp. 584–594, 2008.

[172] M. Miličić Brandt, A. Rath, A. Devereau, and S. Aymé, "Mapping orphanet terminology to umls", in *Artificial Intelligence in Medicine*, M. Peleg, N. Lavrač, and C. Combi, Eds., vol. 6747, Springer Berlin Heidelberg, 2011, ch. 24, pp. 194–203, ISBN: 978-3-642-22217-7.

[173] P. Hougland, J. Nebeker, S. Pickard, M. Van Tuinen, C. Masheter, *et al.*, "Using icd-9-cm codes in hospital claims data to detect adverse events in patient safety surveillance", *Advances in patient safety: New directions and alternative approaches (Vol 1: Assessment)*, 2008.

[174] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, "Dublin core metadata for resource discovery", Tech. Rep., 1998.

[175] IRDiRC, *International rare diseases research consortium - polices & guidelines*, 2013. [Online]. Available: `http://www.irdirc.org/wp-content/uploads/2013/06/IRDiRC_Policies_Longversion_24May2013.pdf` (visited on 05/30/2017).

[176] S. K. Kumar and J. A. Harding, "Ontology mapping using description logic and bridging axioms", *Computers in Industry*, vol. 64, no. 1, pp. 19–28, 2013.

[177] A. A. Sinaci and G. B. L. Erturkmen, "A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains", *Journal of biomedical informatics*, vol. 46, no. 5, pp. 784–794, 2013.

[178] L. A. B. Silva, C. Costa, and J. L. Oliveira, "Semantic search over dicom repositories", in *2014 IEEE International Conference on Healthcare Informatics*, IEEE, Sep. 2014, pp. 238–246, ISBN: 978-1-4799-5701-9.

[179] T. Nunes, D. Campos, S. Matos, and J. L. Oliveira, "Becas: biomedical concept recognition services and visualization.", *Bioinformatics (Oxford, England)*, vol. 29, no. 15, pp. 1915–6, Aug. 2013.

[180] P. Ciccarese, M. Ocana, L. J. Garcia Castro, S. Das, and T. Clark, "An open annotation ontology for science on web 3.0.", *Journal of biomedical semantics*, vol. 2 Suppl 2, no. Suppl 2, S4, Jan. 2011.

[181] L. Ding, J. Michaelis, J. McCusker, and D. L. McGuinness, "Linked provenance data: a semantic web-based approach to interoperable workflow traces", *Future Generation Computer Systems*, vol. 27, no. 6, pp. 797–805, Jun. 2011.

[182] V. Curcin, S. Miles, R. Danger, Y. Chen, R. Bache, *et al.*, "Implementing interoperable provenance in biomedical research", *Future Generation Computer Systems*, vol. 34, pp. 1–16, May 2014.

[183] S. Weibel, "The dublin core: a simple content description model for electronic resources", *Bulletin of the American Society for Information Science and Technology*, vol. 24, no. 1, pp. 9–11, 1997.

[184] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, *et al.*, "Bioportal: ontologies and integrated data resources at the click of a mouse.", *Nucleic acids research*, vol. 37, no. suppl 2, W170–W173, Jul. 2009.

[185] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, *et al.*, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications", *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, Sep. 2010.

[186] P. L. Schuyler, W. T. Hole, M. S. Tuttle, and D. D. Sherertz, "The umls metathesaurus: representing different views of biomedical concepts.", *Bulletin of the Medical Library Association*, vol. 81, no. 2, pp. 217–22, Apr. 1993.

[187] S. Schenk, P. Gearon, and A. Passant, "Sparql 1.1 update", *World Wide Web Consortium*, 2010.

[188] S. Harris, A. Seaborne, and E. Prud'hommeaux, "Sparql 1.1 query language", *W3C Recommendation*, vol. 21, 2013.

[189] C. Forgy, "Rete: a fast algorithm for the many pattern/many object pattern match problem", *Artificial intelligence*, 1982.

[190] M. Zollino, E. Ponzi, G. Gobbi, and G. Neri, "The ring 14 syndrome", *European Journal of Medical Genetics*, vol. 55, no. 5, pp. 374–380, May 2012.

[191] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, *et al.*, "The fair guiding principles for scientific data management and stewardship", *Scientific Data*, vol. 3, p. 160 018, Mar. 2016.

[192] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, *et al.*, "Gene ontology: tool for the unification of biology. the gene ontology consortium.", *Nature genetics*, vol. 25, no. 1, pp. 25–9, May 2000.

[193] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, *et al.*, "The universal protein resource (uniprot)", *Nucleic Acids Research*, vol. 33, no. suppl_1, pp. D154–159, 2005.

[194] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, *et al.*, "The international protein index: an integrated database for proteomics experiments", *Proteomics*, vol. 4, no. 7, pp. 1985–1988, 2004.

[195] C. Pang, A. Sollie, A. Sijtsma, D. Hendriksen, B. Charbon, *et al.*, "Sorta: a system for ontology-based re-coding and technical annotation of biomedical phenotype data", *Database*, vol. 2015, bav089, Sep. 2015.